DOCUMENT RESUME

ED 293 882 TM 011 477

AUTHOR Linn, Robert L.; And Others
TITLE Study Group on Pre-Collegiate Education Quality Indicators. Final Report.
INSTITUTION Arizona State Univ., Tempe.; California Univ., Los Angeles. Center for the Study of Evaluation.; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; Colorado Univ., Boulder.; National Opinion Research Center, Chicago, Ill.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Nov 87
GRANT OERI-G-086-0003
NOTE 129p.
PUB TYPE Collected Works - General (020) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Academic Achievement; Administrative Policy; Administrators; *Data Collection; Educational Assessment; *Educational Quality; Elementary Secondary Education; National Programs; Student Evaluation; *Teacher Effectiveness; Teacher Evaluation
IDENTIFIERS *Pre Collegiate Education Quality Indicators; *Study Groups

ABSTRACT

The Study Group on Pre-Collegiate Education Quality Indicators was formed to determine means of obtaining information on elementary and secondary educational quality within and across states. Two papers: "State-by-State Comparisons of Student Achievement" (Robert L. Linn) and "The Effectiveness of American Education" (Eva L. Baker), along with meeting reports and ancillary material are presented in this document. State and local school administrators encounter public demand for thorough data on the quality of schools, allowing comparisons with data from other states and districts and with their own historical records. The study attempted to: define the content domain of the quality assessment program, relate the definition and score reporting systems to the validity of inferences based on state-by-state comparisons, measure student achievement and teacher quality, and examine the proposed merger of the National Assessment of Educational Progress (NAEP) and the School and Staffing Surveys (SASS). Recommendations include: a complete merger of the questionnaires and samples from the NAEP and SASS should not be attempted in 1990; informing policy analysis should guide any possible merger; a subset of questions from SASS could be administered with the NAEP to enhance policy analysis; and a 3- or 4-year cycle for SASS data collection should be considered. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

C·R·E·S·S·T

Center for Research on Evaluation, Standards, and Student Testing

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Lisa Siegel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED293882

Center for Research on Evaluation, Standards, and Student Testing

Deliverable - November 1987

Study Group on Pre-Collegiate Education Quality Indicators

TM011477

Center for Research on Evaluation,
Standards, and Student Testing

Deliverable - November 1987

Study Group on Pre-Collegiate Education
Quality Indicators

Final Report

Study Directors: Leigh Burstein
Eva Baker

Grant Number: OERI-G-86-0003

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented, or reported herein, was performed pursuant to a Grant from the Office of Educational Research and Improvement/Department of Education (OERI/ED). However, the opinions expressed herein do not necessarily reflect the position or policy of the OERI/ED and no official endorsement by the OERI/ED should be inferred.

State-By-State Comparisons of Student Achievement: The Definition of the Content Domain for Assessment

Robert L. Linn

University of Colorado, Boulder

Twenty years ago when the National Assessment of Educational Progress (NAEP) was being designed, care was taken to ensure that the data would not allow comparisons among individual states or localities. There were a variety of reasons for this decision, including considerations of cost, political viability, and concerns about the likely misuse of state average scores on the assessment. Today, however, the lack of information at the level of individual states has been judged to be the most serious weakness of NAEP by the blue ribbon panel that was constituted to review NAEP and make recommendations about its future (Alexander-James, 1987).

The NAEP Study Group, which was chaired by Governor Lamar Alexander and directed by H. Thomas James, identified the development of state-by-state comparative data as its number one priority. The Study Group reasoned that most "important decisions in education are made at the state or local level, and accountability for performance is vested at those levels" (p. 4). They also implied that the decision makers at the state or local level would benefit from comparative information, but did not explicitly state how such information would be used to make better educational decisions.

The Study Group considered some of the concerns that, in the past, had led to a decision to prevent the use of NAEP for purposes of making state-by-state comparisons, but concluded that the "concerns are less important now than they were previously, and that most can be readily accommodated within a redesigned national assessment" (Alexander-James, 1987, p. 5). Having thus dismissed the objections to state-to-state comparisons under the heading "previous concerns

about comparisons", the Study Group was ready to give its most important recommendation.

> The single most important change recommended by the Study Group is that the assessment collect representative data on achievement in each of the fifty states and the District of Columbia. Today state and local school administrators are encountering a rising public demand for thorough information on the quality of their schools, allowing comparison with data from other states and districts and with their own historical records. Responding to calls for greater accountability and for substantive school improvements, state officials have increasingly turned to the national assessment for assistance (pp. 11-12).

The movement toward state-by-state comparisons, of course, did not begin with the Alexander-James Study Group. Rather, the Study Group endorsed a position that had already garnered considerable support from policy makers during the past five years and pointed to the redesign of NAEP as a mechanism for obtaining the desired comparisons. The movement toward state-by-state comparisons was encouraged earlier by the U.S. Department of Education and by the Council of Chief State School Officers.

The Council of Chief State School Officers has provided considerable support for the idea of state-by-state comparisons during the past three years since the Council adopted a position paper encouraging states to develop comparable measures of student achievement in reading, mathematics, English, science, and social studies. The subsequent establishment of the State Education Assessment Center by the Council with the support of the Center for Statistics and the Mott Foundation and the activities of the Assessment Center and the Council since that time have given greater strength to the movement toward making state-by-state comparisons a reality. With support from the U.S. Department of Education and the National Science Foundation, the Council is now in process of forming a consortium of educators that will develop specific recommendations for the first state-by-state assessment of student achievement in mathematics.

As Ramsay Selden (1986a), the director of the State Education Assessment Center, has noted, any approach that is taken to the development of a system that will yield state-by-state comparisons of student achievement will raise "profound issues in educational measurement" (p. 2). Selden went on to discuss some of those issues and highlighted the need to deal with issues of validity. The focus of this paper is on a limited set of issues related to the validity of the assessment system. More specifically, the purpose of this paper is to review issues concerning the definition of the domain of content to be covered in the assessment and the relationship of the definition and score reporting systems to the validity of inferences that are based on state-by-state comparisons.

Validity

As with any use of tests, the most fundamental measurement issue in the development of an assessment system that will provide state-by-state comparisons is the validity of the inferences that will be made from the scores. To date, however, relatively little serious attention has been given to the questions of validity of a NAEP based state-by-state comparison system, or for that matter, any other system other than the seriously flawed use of ACT and SAT scores as indicators of the educational quality in a state.

Although not couched in terms of validity, the primary concern that was raised in the National Academy of Education's review committee commentary on the Alexander-James report is fundamentally an issue of validity. The Review Committee (National Academy of Education, 1987, p. 59) summarized its reservations about the recommendation that NAEP be redesigned to provide state-by-state comparisons as follows.

> We are concerned about the emphasis in the Alexander-James report on state-by-state comparisons of average test scores. Many factors influence the relative rankings of states, districts, and schools. Simple comparisons are ripe for abuse and are unlikely to inform meaningful school improvement efforts.

As is clearly implied by the above statement, the Review Committee's concern applies not only to the proposed state-by-state comparisons using NAEP but to the use of average test scores for other units such as individual school buildings or school districts. The concern is not limited to the use of NAEP. It would apply equally well to the use of other assessment devices or tests. The concern is clearly with the inferences that the Review Committee anticipated would be made from the test data and the validity of those inferences will depend on a wide variety of factors, such as the degree of standardization of the rules for inclusion and exclusion of students in the assessment, the specific sampling procedures, and the administration procedures. One of the important factors that will influence the validity of the inferences drawn from the comparisons, however, is the adequacy of the content coverage of the assessment.

Content Domain

It is one thing to agree that the assessment should cover the "core content areas (reading, writing, and literacy; mathematics, science, and technology; history, geography, and civics)" (Alexander-James, 1987, p. 12), but quite another to agree that a particular set of topics in, say, history, much less that a specific set of items, should be included on the assessment that is to be used to compare states. It is also much easier to achieve agreement that "the assessment instruments should examine acquisition of pertinent 'higher-order' skills as well as basic skills, knowledge, and concepts" (Alexander-James, 1987, p. 8), than it is to gain consensus that a given exercise is a fair assessment of higher-order thinking skills. Many of the issues that arise when a school or district selects a test are also relevant at the state level. Among these are the issues of the breadth of the coverage, the match between what is taught and what is tested, the number and specificity of the scores that are reported, and

the familiarity of the assessment procedures that are used.

Breadth of Coverage and the Match with What is Taught. Since the issues of breadth of coverage and that of the degree to which the assessment matches the curriculum and what is actually taught in classroom are closely related, they will be considered together. One approach to the determination of the content to be included in an assessment would be to require a consensus among all states that a given topic or assessment exercise is appropriate to the state's instructional goals for students at a given point in their educational program. As Selden (1986a) has noted, the consensus about a "common body of knowledge could be conceived as a 'least common set' -- that content which is pursued to some degree by schools in each [state], but excluding anything which all states cannot be presumed to be teaching or emphasizing. Alternatively, it could be conceived as an 'optimal set', around which consensus can be reached, but which may not reflect everything some states are pursuing, and which may include some items that some states may not be pursuing or emphasizing" (p. 7). To these two alternatives could be added, at least in theory, an "inclusive set", that content that is judged to be appropriate by one or more states.

Although the "inclusive set" is apt to be too unwieldy in practice, it illustrates an end of a continuum that is anchored at the other end by the "least common set". On the surface the least common set appears the fairest approach. It would not hold a state accountable for students learning content that was not expected to be taught in its schools by a given grade level. However, as will be discussed in some detail below, the least common set approach can be faulted on several accounts including that of fairness.

The issue of where along the continuum between the least common and the inclusive sets an assessment should be placed is not unique to the present context. It long has been an important issue in the use of tests in program and curriculum evaluation (e.g., Burstein, 1981; Cronbach, 1963; Walker &

Schaffarick, 1974; Wargo & Green, 1978). If a test does not measure the outcomes that correspond to important program goals, the evaluation will surely be considered unfair. The judgment that the evaluation is unfair takes on additional force when multiple programs are compared and the tests used to measure the educational outcomes of the programs appear to match the goals of one program better than another.

The latter point is clearly illustrated by the controversy that surrounded the Follow Through evaluation. Follow Through was a massive federal experiment that pitted twenty-two early education models against each other over the course of ten years. The model programs varied considerably in their stated goals but were evaluated using a common set of outcome measures. Between-model differences were found on some of the subtests of the Metropolitan Achievement Test (MAT) (Stebbins, St.Pierre, Proper, Anderson, & Cerva, 1977). The differences occurred on subtests that the evaluators classified as "basic skills" and favored models that were classified as emphasizing basic skills over models that were classified as having a "cognitive-conceptual" emphasis or an "affective-cognitive" emphasis. Press accounts of the evaluation presented the message that education that emphasizes the basics yields the best results.

Because of the potential importance of the Follow Through evaluation, the Ford Foundation sponsored a comprehensive third-party review of the evaluation. The review resulted in a devastating critique that faulted the evaluation on numerous grounds (House, Glass, McLean, & Walker, 1978). Of most relevance to the present discussion, however, is the House, et al. critique of the measurement of the program outcomes and the characterization of those outcomes. Their analysis led them to conclude that "the outcome measures assess very few of the models' goals and strongly favor models that concentrate on teaching mechanical skills" (House, et al, 1978, p. 156).

Although not strictly a question of test content, the format of the test items and administrative procedures can also have implications for the results of an assessment. Even apparently trivial changes in item format, such as the presentation of addition problems horizontally rather than vertically, have been found to effect the scores that children obtain (Alderman, Swinton, & Braswell, 1979). More importantly, the outcome of an assessment can be affected by the match between the format used to ask question on the test and the format used when students practice the skill in the instructional program and the amount of practice that they have with similar tests (Alderman, et al., 1979; Cooley & Leinhardt, 1980; Roberts, 1980).

The match between what is taught and what is tested can have a substantial effect on the performance on tests. The closer the match and the more the test questions tap rote memory, the larger the likely effects. Indeed, two of the most compelling examples involve the choice of words for tests of spelling or for the vocabulary used to assess beginning reading. Hopkins and Wilkerson (1965) compared four forms of the California Spelling Test to the course of study guide used in California. Because the forms varied in the degree to which they matched the study guide, knowledge of only those words that were in the curriculum study guide would yield scores that differed by as much as 2.1 grade equivalent units depending on which of the four forms was used. As would be expected, the California students were much more likely to correctly respond to words that were in the curriculum than words that were not.

Bianchini's (1978) analysis of the remarkable increase in the percentile rank of the median reading achievement test score for first grade students between 1970 and 1971 provides another example of the dramatic effect that the degree of match between what is taught and what is tested can have on tests scores. Over the course of that single year, the median score for first grade students throughout the state rose from the 38th to the 50th percentile. As

Bianchini's analyses suggests, however, the huge increase had more to do with the fact that the test that was used to measure reading achievement was different in 1971 than it was in 1970, than to any dramatic increase in the quality of education provided to first grade children. Bianchini found that 55% of the vocabulary on the test that was used in 1971 was included in the state's first grade readers, whereas only 19% of the vocabulary on the test used the previous year was included in the readers.

Results such as those reported by Bianchini (1978), Hopkins and Wilkerson (1965), and others (e.g., Cooley & Leinhardt, 1980; Leinhardt, 1983; Leinhardt & Seewald, 1981) might lead one to believe that "least common" set approach is necessary to avoid unfair comparisons. However, the solution is not that simple. To begin with, the fact that two programs both teach children to add two digit integers, for example, does not imply that both programs give that skill the same priority or spend an equal amount of time teaching it. If the children at one school were drilled extensively on the addition of two digit integers, with little attention given to other arithmetic operations or to mathematics concepts, while children at a second school spent some, but much less, time on that skill while spending considerably more on other skills and on concepts and problem solving, a test that only measured the addition of two digit numbers would hardly be considered fair. As in the case of the Follow Through evaluation, the test would strongly favor the first school because it lacked more comprehensive coverage of the skills and concepts that were emphasized at the second school. While such extremes are unlikely to be encountered in practice, even at the level of individual schools much less at the level of entire states, the example illustrates the fact that the use of the least common set will tend to favor those who emphasize the skills and concepts contained in that set at the expense of those that are not included in the set.

No matter what process is used to define the domain of content, it must include knowledge, skills, and concepts that educators, policy makers, and the general public consider important. This is part of the reason that the Alexander-James (1987) report emphasized that the assessment should include measures of higher order skills which the report defined to include "recognizing a problem's general structure, defining goals, isolating the information relevant to problem solutions, ... evaluating the merits of arguments, ... reasoning, analyzing, explaining, and finding analogies" (p. 15). Such a list does not appear to be compatible with the least common set approach to defining the domain to be assessed for purposes of state-by-state comparisons.

The Alexander-James (1987) list of higher-order thinking skills would push the assessment beyond a minimum set of basic skills that would be likely to define the least common set to a broader set of goals. In as much as there is general agreement that higher-order thinking skills of the type envisioned by the Alexander-James study group should be taught, the list is in keeping with what Selden (1986b) has referred to as the "optimal consensus" approach wherein the content of the assessment would be defined to include content for which a consensus can be reached that given content knowledge and skills should be taught. The idea of this approach is that it would allow the assessment to go beyond minimal objectives that are already pursued by all and thereby have a aotentially broadening influence on the curriculum rather than a narrowing influence that is apt to be associated with least-common-set approach.

If the assessment is to encourage greater breadth and depth of content coverage, it will need to have a content domain with broadly defined limits and emphasize more than simple factual knowledge. As Anderson (1986) has note, such an assessment is apt to measure several dimensions of achievement within each subject area and raise questions about the nature and number of scores to be reported.

Number and Specificity of Scores. Cronbach (e.g., 1963, 1971) has long argued that for purposes of evaluation, a comprehensive array of measures should be sought. "An ideal evaluation might include measures of all the types of proficiency that might reasonably be desired in the area in question, not just the selected outcomes to which ... [a particular] curriculum directs substantial attention" (Cronbach, 1963, p. 680). The assessment needs to provide a basis for identifying areas that are judged to be important but that students are not learning, whether or not the poor learning is the result of lack of exposure. Furthermore, for purposes of making decisions about the curriculum or program of instruction, the test results need to be reported separately for each of the specific areas of proficiency, and not merely combined into a single overall score.

The latter point runs counter to the goal of having a simple scorecard that will allow the ranking of states along a single dimension. However, Cronbach's rationale for maintaining separate scores is compelling.

> If the original test or battery is a composite covering various types of content or various objectives, it implicitly weights those elements, either by the number of items allocated to each or by the way the score is calculated. Such a weighting cannot satisfy decision makers who hold values unlike those of the test developer. Consequently, an ideally suitable battery for evaluation purposes will include separate measures of all outcomes the users of the information consider important ... Reporting separate scores allows for the application of various systems of values. It also enables the investigator to examine the nature of any weaknesses in the program. (emphasis in the original) (Cronbach, 1971, p. 460).

The use of a single composite score not only forces an implicit set of values on the outcome of the assessment and prevents those who hold different values from seeing the results from that alternate framework, but the composite may sometimes be insensitive to differences between the educational systems that are being compared (Airasian & Madaus, 1983; Madaus, Airasian, & Kellaghan, 1980). In other instances, and of even greater concern, the composite may favor a system with an emphasis that happens to match the content that the composite

weights most heavily.

The latter problem is illustrated by the results of Walker and Schaffarzick's (1974) review of twenty-six studies that compared students who had been exposed to a given subject matter using either "traditional" or "innovative" curriculum materials and then tested with one or more measures of achievement. Their review provides strong evidence that "different curricula are associated with different patterns of achievement" (emphasis in the original) (p. 97). Whether the results of the studies reviewed favored the "traditional" or the "innovative" curriculum was largely determined by the content of the tests. "Students using each curriculum do better than their fellow students on tests which include items not covered at all in the other curriculum or given less emphasis there" (Walker & Schaffarzick, 1974, p. 97). If a single global score were used to compare the alternative curricula an outcome of no difference, one favoring the traditional curriculum, or one favoring the innovative curriculum could be readily achieved according to the relative weighting given to the test content favoring each.

The need to report multiple scores corresponding to narrowly defined content areas is clearly demonstrated by recent experience with tests that are customized to the specifications of a state or local district. The need for multiple scores can also be demonstrated from recent experience with the NAEP assessments in literature and U. S. history. In both instances it is evident that a single total score can conceal specific areas of strength and weakness. Furthermore, the relative standing of a given state, region, or other aggregate of students can be greatly influenced by the number of items that happen to be associated with specific content areas.

In the past, if a state or district wanted to compare the achievement of its students to a national norm, it had to administer a norm-referenced test.

If the state or district also wanted to obtain test results on a test designed to match locally defined objectives, a second test administration was generally required since the standardized test would not match the locally defined objectives as closely as desired. Recently, however, test publishers have begun offering an option of creating a "customized" test that consists of items selected according to locally specified objectives, but from which norm-referenced scores are also produced.

Customized tests are the result of increased use of item response theory by publishers in their test development and scaling process. One of the features of item response theory that makes it especially appealing is the promise that, once the theory has been used to calibrate a pool of test items, any set of items from that pool can be used to place the performance of test takers on a common scale. Thus, according to the theory, any set of previously calibrated items could be selected by a state or district to be included among those on its customized test and the resulting test scores could still be placed on the same scale as the published version of the standardized test for which national norms are available.

The quality of the norm-referenced scores that a state or district obtains for its customized test depends on several factors, including (1) the adequacy of the item response theory model for the set of items in the calibrated item pool, (2) the number of calibrated items selected for the customized test, (3) the statistical characteristics of the items selected from the item pool, and (4) the degree to which items selected for the customized test match the content coverage of the published version of the test for which the norms are available. Recent experience with a major customized test, the Kentucky Essential Skills Test (KEST), suggests that the last of these four considerations can be of critical importance (Linn, 1986; Yen, Green, & Burkett, 1987).

Kentucky administered the KEST to essentially all eligible students in the

state in grades K through 12 for the first time in 1985. The 1985 KEST was a customized test, containing, among other items, items that were selected from the CTB/McGraw-Hill item pool. That pool includes items from the Comprehensive Tests of Basic Skills (CTBS), Forms U and V, items from the California Achievement Tests, Forms C and D, and previously unpublished items. Since all items are calibrated to the CTBS scale, a test that had previously been administered statewide in Kentucky, it was possible to obtain estimates of performance on the CTBS scale from the administration of the KEST. When the KEST results were obtained in 1985, however, at least two major anomalies were observed. The the most notable and troublesome of these was a precipitous increase in the grade 5 mathematics test performance.

In 1982, 1983, and 1984, when the CTBS was administered statewide to fifth grade students, the state mean normal curve equivalent (NCE) scores in mathematics ranged from 50.4 to 54.8. In 1985, however, the mean NCE for grade 5 mathematics based on the KEST was 66.3. Thus, on the NCE scale, which has a standard deviation of 21 for the national norm group, the state mean increased in a single year by slightly over a half of the national norm group standard deviation. Although a review of the KEST and the calibration of the items in the item pool from which it was constructed did not suggest that the application of item response theory was any more problematic than in many other widely accepted applications, it was evident that the grade 5 mathematics results on the 1985 KEST could not be meaningfully compared to the earlier CTBS results (Linn, 1986).

The lack of comparability between the KEST and CTBS grade 5 mathematics tests is most plausibly explained by differences in the proportion of items on the KEST and the CTBS that are classified into specific content categories. The proportions of KEST and CTBS grade 5 mathematics items by content category

were as follows (Linn, 1986).

Content Category	CTBS Proportion	KEST Proportion
Numeration	.42	.27
Number Theory	.03	.13
Measurement	.16	.11
Geometry	.10	.20
Number Sentences	.19	.20
Problem Solving	.10	.09

As was demonstrated by Yen, Green, and Burkett (1987), systematic differences as a function of content category between local and national estimates of item response theory difficulty parameters are sometimes found. Such differences can lead to misleading global score results when content coverage changes. "Content equivalence between customized and normed tests is essential if the customized test is to be NRT-equivalent and norm-valid" (Yen, Green, & Burkett, 1987, p. 13). Separate reporting by specific content categories, however, is needed in order to identify areas of strong and weak performance and to make value judgments about the importance of changes in scores on the global score.

The final example illustrating the importance of multiple scores corresponding to specific content categories comes from the recent NAEP results in literature and U. S. history (Applebee, Langer, & Mullis, 1987). Both the literature and the U. S. history item sets met the usual criteria for deciding if a unidimensional item response theory model is appropriate. Hence, single global performance scores were estimated for each of the two broad content domains.

Despite the apparent simplicity for each content area, however, substantial differences that could be meaningfully interpreted were found for content specific subsets of items as a function of region of the country, gender, and race/ethnicity. For example, even though the performance of black test takers

was well below that of whites on the bulk of the literature and U. S. history items, blacks outperformed whites on questions asking about black leaders or black literature. Black test takers also did better than whites on several of the questions dealing with slavery and civil rights. Similarly, though women outperformed men on the overall literature scale, men did better on "items focusing on strong male literary characters" (Applebee, et al., 1987, p. 3), such as Robin Hood, King Arthur, Samson, and Captain Ahab. Although the Southeast region of the country scored well below the northeast on the overall literature scale, the converse was true on the 15 items dealing with Biblical characters and stories.

The above examples illustrate two points that are of great potential importance in any future state-by-state comparisons of student achievement. First, the rank order on a single global score is apt to depend on the particular weighting of the content categories. Based on the KEST results, one might reasonably expect, for example, that Kentucky would have appeared better on a grade 5 test with heavy emphasis on numeration than on one that emphasized another content category such as number theory or geometry. Second, a single global score can also conceal educationally important information about strengths and weaknesses in the curriculum.

The need to focus on multiple content specific outcomes has been recognized within the context of state assessments by Bock and his colleagues (Bock & Mislevy, 1987; Bock, Mislevy, & Woodson, 1982; Mislevy, 1983). For purposes of informing curriculum planners, assessment information needs to be provided for highly specific content areas which Bock, Mislevy, and Woodson (1982) called "indivisible curricular elements". These are "item domains that are sufficiently homogeneous with respect to content that all the items in a given domain would be similarly affected by changes in curricular emphasis" (Mislevy, 1983, p. 273).

Summary and Conclusions

It has been argued that the choice of content for a state-by-state will be one of many factors that will have a substantial influence on the validity of inferences that may be drawn from a state-by-state assessment system. Based on considerable experience in the use of tests in the evaluation of alternative educational programs, it was concluded that there are great disadvantages to an approach that focuses only on content and skills that are thought to be taught by a given grade in all states. Such a "least-common set" approach would be likely to give a relative advantage to states that narrow their focus to only that least common set. The approach is more likely to narrow than to broaden the curriculum.

Ideally, the domain for assessment would include separate measures of the full range of outcomes that are considered important by any of the states. The multiple measures would enable states to identify strengths and weaknesses and not just obtain a ranking on a global scorecard. The more inclusive set would encourage a broadening rather than a narrowing of the curriculum by calling attention to wide range of outcomes.

Despite the desirability of having multiple scores corresponding "indivisible curricular elements" for purposes of identifying strengths and weaknesses and planning changes in the curriculum, such scores clearly will not satisfy the demand for a overall number in reading or a single score for mathematics. Global scores will certainly need to be produced, in part, because the amount of information would be too overwhelming for many of its intended uses if it were only reported at the level of indivisible curricular element level, and, in part, because there is a desire, as Ambach (1987) has noted, for a scorecard. Global scores can, and undoubtedly, will be produced. The argument here is not that such scores should not be produced, but that the

ability to disagregate the results to more specific content areas should be maintained. The disaggregated scores are needed to interpret the overall results and plan improvement.

References

Airasian, P. W. & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. Journal of Educational Measurement, 20, 103-118.

Alderman, E. L., Swinton, S. S., & Braswell, J. S (1979). Assessing basic arithmetic skills and understanding across curricula: Computer-assisted instructional and compensatory education. Journal of Children's Mathematical Behavior, 2, 3-28.

Alexander-James Study Group. (1987). The nation's report card: Improving the assessment of student achievement. Cambridge, MA: National Academy of Education.

Ambach, G. M. (1987). Testing and educational quality. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1985) Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Anderson, P. S. (1986). Beyond the wall chart: Issues for states. Technical Report. Portland, OR: Northwest Regional Education Laboratory.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1987). Literature and U. S. history: The instructional experience and factual knowledge of high school juniors. Princeton, NJ: Educational Testing Service.

Bianchini, J. D. (1978). Achievement tests and differential norms. In M. J. Wargo & D. R. Green (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill.

Bock, R. D. & Mislevy, R. J. (1986). Comprehensive educational assessment for the states: The duplex design. Technical Report, Chicago. IL: National Opinion Research Corporation.

Bock, R. D., Mislevy, R. J., & Woodson, C. E. M. (1982). The next stage of educational assessment, Educational Researcher, 11, 4-11, 16.

Burstein, L. (1981). Investigating social programs when individuals belong to a variety of groups over time. CSE Technical Report # 173. Los Angeles, CA: UCLA Center for the Study of Evaluation.

Cooley, W. W. & Leinhardt, G. (1980). The instructional dimensions study. Educational Evaluation and Policy Analysis, 2, 7-25.

Cronbach, L. J. (1963). Evaluation of course improvement. Teachers College Record, 64, 672-683.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement, Second edition. Washington, DC: American Council on Education. pp 443-507.

Hopkins, K. D. & Wilkerson, C. J. (1965). Differential content validity: The California Spelling Test. Educational and Psychological Measurement, 25, 413-419.

House, E. R., Glass, G. V., McLean, L. D., & Walker, D. F. (1978). No simple answer: Critique of the Follow Through evaluation. Harvard Educational Review, 48, 128-160.

Leinhardt, G. (1983). Overlap: Testing whether it's taught. In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing. Hingham, MA: Kluwer Nijhoff.

Leinhardt, G. & Seewald, A. (1981). Overlap: What's tested, what's taught? Journal of Educational Measurement, 18, 85-96.

Linn, R. L. (1986). Norm-referenced score estimates from the Kentucky Essential Skills Test. In Center for the Study of Testing, Evaluation, and Educational Policy (George Madaus, Principal Investigator). An evaluation of the Kentucky Essential Skills Test. Chestnut Hill, MA: Boston College.

Madaus, G. F., Airasian, P. W., & Kellaghan, T. (1980). School effectiveness: A reassessment of the evidence. New York: McGraw-Hill.

Mislevy, R. (1983). Item response models for grouped data. Journal of Educational Measurement, 8, 271-288.

National Academy of Education. (1987). Review Committee commentary of The Nation's Report Card. Cambridge, MA: National Academy of Education.

Roberts, A. O. H. (1980). Practice effect of test-wiseness. Mountain View, CA: RMC Research Corporation.

Selden, R. (1986a). Some classical measurement issues confronting the development of state-by-state assessment of student achievement. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA, April.

Selden, R. (1986b). White paper. Strategies and issues in the development of comparable indicators for measuring student achievement. State Education Assessment Center, Council of Chief State School Officers., April 30.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B., & Cerva, T. R. (1977). Educational as experimentation: A planned variation model, Volume IV-A, An evaluation of Follow Through. Cambridge, MA: Abt Associates, Inc.

Wargo, M. J. & Green, D. R., (Eds.). (1978). Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill.

Yen, W., Green, D. R., & Burkett, G. R. (1987). Valid normative information from customized achievement tests. Educational Measurement: Issues and Practice, 6, 7-13.

The Effectiveness of American Education

Eva L. Baker
UCLA Center for the Study of Evaluation

American political attention has turned with increasing intensity to the matter of educational quality. From the reports of commissions and panels to debates by presidential candidates, the focus on students, teachers, and schools grows sharper every day. At the center of concern is a deceptively simple question: How well do our schools prepare our students?

It doesn't matter if the language emphasizes excellence, subject matter understanding, productivity, or competitiveness, the meaning of the debate is clear: Can we describe, judge and improve the effectiveness of public schools?

Over the years, significant investments have been made in trying to answer these questions. Standardized achievement tests, educational program evaluations, teacher testing, and minimum competency tests for students all are thought to provide useful information to help make judgments about the effects of educational services on students. Many of these options have roots in the mid-sixties enactment of federal legislation to assist educationally disadvantaged students. This new legislation required that the federal government evaluate the effects of its efforts to provide compensatory resources for students. The legislation was directly responsible for the rapid development and growth of the evaluation field and for many scientific developments in the measurement of human performance. Through the ensuing decades, one or another particular version of evaluation or measurement was selected as the new solution for understanding school effectiveness, the options coming, it seemed, in overlapping waves. Remember? Different solutions included setting objectives and measuring student

performance, local standardized student testing, program evaluation, Scholastic Aptitude Examination (SAT) score decline, state minimum competency examinations, teacher testing, state assessment, and "The Wall Chart," a national comparison of educational systems. None of these approaches were found to be wholly satisfactory, but, after the initial blaze of interest died down, none were retired either. Instead, our attempts to understand educational quality have resulted in an increasing set of measures and approaches designed to shed some light on the issue. But do they? Imagine that we could start over, fresh and unsullied by our prior measurement experience. What would be fair measures of the effectiveness of our educational programs?

To answer this question, we first must decide what level of information we want. Making a judgment about all of American education and assessing the effectiveness of First Street School in your hometown require different levels of information. In the first case, we would look for common features of schools and curricula to base our judgment. When looking at a particular school, however, we can be much more attentive to the community characteristics, the kinds of students attending the school, the particular goals of the school, and other special conditions. In both cases however, we simply want to know the following:

What are the students learning?
How well do the teachers teach?
What is the quality of our schools?

The public seems equally interested in the concrete accomplishments of local schools and the general descriptions of the educational system at large.

Educators want answers to these questions. These answers should not simply describe the state of performance for students, teachers, and school administrators, but should ideally permit us to devise actions to make things better. We want information for more

than curiosity's sake; we want it to help us improve education. This desire to face and fix what's wrong requires that the information we collect gives us more than categorical "good" or "poor" labels. We need enough detail to guide our policies and practices.

With this discussion as preface, let's consider in turn questions of effectiveness that involve students, teachers, and schools.

Student Learning

Student learning has been traditionally measured by achievement tests. For public accountability purposes, teacher-made tests have never been regarded as sufficient. Rather, because accountability implies some sort of comparison, tests that have standard content and rather general applicability have been used. Without rehashing two decades of concerns about standardized testing, a few issues remain salient:

- Standardized tests allow comparisons among schools and regions. They may, however, be somewhat insensitive to curricular and instructional variations. Because they are prepared to be of widest utility, standardized tests may omit areas of particular emphasis for particular schools. These tests provide information on only a narrow slice of school activities.

- Standardized tests most often ask children to answer questions given in multiple-choice format. I believe this format greatly underestimates student performance.

- Because of technical reasons used in test statistics, very small absolute differences (for instance, one test item) might mean an improvement of a "grade level" or so. Making inferences about educational quality based on these differences is a shaky proposition.

Test performance still is, in that unfortunate phrase, the bottom line for many who would assess the effectiveness of the schools. At this time, standardized tests are regarded by many policymakers as credible and objective. Achievement testing will not go away, and for good reason. Students and, by implication, the schools to which they go must be held accountable for teaching students and for attempting to measure what they have learned. Standardized tests are thought by many to be the best approach we have.

But these tests can be greatly improved. At the Center for Research on Evaluation, Standards, and Student Testing (CRESST), sponsored by the US Office of Educational Research and Improvement, we are in the midst of a five year research program to improve the quality of testing for use in the schools.

The precepts of our program, and the way we believe testing ought to be improved, fix on a small set of critical issues. In one way or another, our attention focuses on validity, or the quality of the information the test provides us and the degree to which we can believe it.

Validity. Validity of achievement measures has a number of components (See Baker and Herman, 1986, for a fuller discussion). One critical component is the degree to which the way performance is measured matches the mode in which learning best occurs. With the advances in cognitive science, we believe we can design measures that more productively represent the richness of learning. For example, we are interested in assuring that in mathematics, science, and history, students be given different ways to demonstrate their competence, perhaps in multiple choice tests, perhaps in other paper and pencil formats, perhaps using computer dynamic displays, perhaps in writing. Many current testing formats developed out of convenience for the administration and scoring processes rather than

because they were the best ways to assess complex human understanding. One attribute of tests is that they often force students to give the first, quick response, rather than a thoughtful, reasoned answer. The balance between conserving the time spent on testing and providing enough opportunity for adequate thought is still unsettled. Perhaps a more diverse menu of testing approaches will increase the overall validity of our measures, and allow testing approaches to match better student propensities.

A second validity concern relates to the content or subject matter of what is to be tested. One of the sadder outcomes of the behavioral objective movement and of inquiry approaches of the early seventies was the attention paid to process *at the expense* of the content to which these processes applied. We have seen the pendulum swing widely on this issue during the last two decades. Given the popularity of books like *Cultural Literacy* (Hirsch, 1987) and the scandalous blanks and misunderstandings in our students' knowledge, we are again on the verge of another swing towards content. It's tempting to devise tests that can pinpoint such content errors. This time, however, we want to assure that we go well beyond identification or recognition of specific facts and concepts. We intend to integrate measurement approaches that wed content with sophisticated approaches to demonstrating understanding, such as complex essays. We at UCLA are developing the technology to score such essays reliably and relatively inexpensively.

Third, we are interested in measures that can be related directly to instructional options. We should be measuring performance that schools can affect. This means that, where possible, we should be collecting information about teaching practices, student familiarity with content, and so on, at the same time that we measure student performance.

Fourth, our measures must be valid when individual and group difference are considered. Whether a test is fair is a psychological as well as an empirical issue. We particularly want to assure that our

measures validly assess strengths and weaknesses of our pluralistic student body in a way that contributes to their motivation to continue learning.

Quality of interpretation. Even when student achievement is measured validly, the way such findings are interpreted makes a difference. Interpretation involves relating findings to other similar measures of performance, comparing findings to the performance of other similar groups of students or schools, analyzing findings in the light of previous performance to see the development of trends over time, or looking at performance in terms of some predetermined standard. Comparison to other student groups is the most common interpretation strategy. This comparison is the basis of "norms," or averages provided for many nationally standardized tests. In some state assessments, comparisons for student performance are provided by looking at the performance of students in schools of similar size and community location. More recently, the federal government has reported the comparison of student performance on the SAT state by state, a specific approach to be discussed later.

A central issue of interpretation is what is being compared. Are tests of individual students used to make comparisons among schools? What other information needs to be collected if such use of information is to be sensible?

The first question for these sorts of comparisons is: "Is the comparison fair?" One shouldn't compare a small, stable suburban school with a central city school that has a high mobility rate. Given the increasing diversity of our students, comparisons now must involve issues such as language in the home and length of time in the school in addition to the more usual socioeconomic measures.

Other options have been the international comparisons, where we look at US students in comparison to those in other countries. While such comparisons might be useful is setting goals for our students, the inference remains that we should adopt practices

embedded in other cultures or in other constitutional, and more centralized, arrangements for education policymaking. Such an inference is probably unwarranted.

Moreover, the bane of most normative comparisons is that half of the group is always below average, a status unacceptable to most educational policymakers. No one yet has figured out how all students can perform "above the average."

To sum up, what should we want in student achievement measurement?

- More than one measure of the same phenomenon, such as reading comprehension (to allow for corroboration from different sources), but with no expectation that all students need to take multiple measures.

- More than one kind of testing format, such as multiple choice and written answers.

- Tests that give students adequate time to perform serious cognitive tasks.

- Tests that measure both the content (what) and the process (how) that students use to solve complex problems of understanding.

- Tests that can be analyzed to guide instructional planning.

- Test results that are understandable, timely, and usable by teachers for instruction and planning (see Herman and Dorr-Bremme, in press, for a report of teachers' test use.

- Reports of test results that are fair to students, teachers, and schools.

Quality of Teaching

A second enduring concern in education is the quality of teaching. This interest is obvious; when we think of schools we think of teachers. Given the instructional and economical dominance of teachers in schooling, it's natural to want to judge effectiveness of educational investments in part by looking at teaching. The problems begin when one tries to operationalize the measurement of the quality of teaching and confuses it with the "quality" of teachers. Just as in the student achievement area, the principal trouble spot in quality of teaching is validity. There is little real agreement on what good teaching is. When good results occur, we can attempt to infer which teaching practices were responsible. A general application of principles such as providing students with opportunity to learn, clear task directions, and feedback, undoubtedly apply on the average. Our problem is that we are often not interested in teaching on the average, but are particularly interested in a particular teacher's competency, perhaps for merit pay or other forms of advancement. When the individual teacher is our focus, we must take special care to allow adequate flexibility in pedagogical style, since for various topics, objectives, grade levels, personalities, settings, and student groups, no "best" pedagogical approach has been identified. With support of the Carnegie Foundation, new approaches to the assessment of teaching competencies are under development. Although designed to permit special certification of teachers rather than the assessment of educational effects, their efforts may have some positive influence on the measurement of teaching capability.

Teacher testing. Because teaching quality has been hard to measure, many have supported the measurement of prerequisites that good teachers are presumed to need. Such prerequisites include mastery of subject matter, mastery of basic knowledge about teaching, student development, and learning, and mastery of basic skills. Tests have been devised to assess teachers in many of these areas, some with associated sanctions. Without disputing the right of

the state or school district to set standards of this sort, conflicts have developed on a number of points. Rudner (1987) points out that the standards for many of these tests have been set very low. Lorrie Shepard, in a case study of the Texas Teacher Test (1987) describes how it might be possible to pass the test by being testwise rather than being skilled in the area the test was assessing. Ellwein and Glass (1986) infer from their case study that teacher testing is mostly symbolic and has very little to do with actually identifying deficiencies and improving instruction. Involved in many of the analyses of teacher testing is the question of when it should occur (pre-service? pre-teacher education program?) and to whom the sanctions should apply (the teacher? the degree-granting institution? the teacher training institution?).

Student achievement as a measure of teaching. Using student achievement as a way to estimate teaching effectiveness is another approach. It seems like a reasonable tactic; after all, teachers ought to help students learn. Clearly subject to the validity concerns about student testing listed above, the use of such measures to assess teachers unfortunately adds new complexity. Minimally, these comparisons may necessitate complex tracking of students who enter particular teachers' classes. Statistically equating students with different entry competencies is sure to be an unsatisfactory way to compare teachers' relative merit in promoting achievement. On the one hand, it's harder to teach students who have inadequate backgrounds. Alternatively, it's also difficult (because of artifacts of tests) to show real improvement when the student group comes in with a very strong achievement levels. In either case, the achievement tests will probably misrepresent the nature of the teacher's effort. Thankfully, recent assessment systems for teachers are attempting to represent more broadly the nature of teachers' efforts.

Educational Quality of Schools

Who wants to know? The desire to find out how schools are doing is clearly legitimate, and educators, policymakers, and researchers continue to propose alternative sources of information. One of the problems we face is providing the right information to the right people. Congressional policymakers want to know whether the schools are working (Congress of the United States Congressional Budget Office, 1987). At different times, their concern may be focused on the quality of what is learned (as in the post-Sputnik period) or who is learning (when equity concerns are central on the educational agenda). Their needs are to assess the impact of resources they have invested and to target continuing or new needs. They need relatively unambiguous, clear information. To even a greater extent, state level policymakers are concerned with the effects of specific policies related to financing, curriculum, and certification, i.e., their efforts to reform schools in their states. Local school boards and their administrations have needs for information related to the quality of their policy implementation and the progress toward discretionary goals, given the particular characteristics of their community. Each set of policymakers has differential need for detail and different opportunity to influence the reality of classroom practice. The hodgepodge of conflicting information from local, state, and national evaluations doesn't make evaluation of educational effectiveness any easier. Some new approaches may offer some relief.

Comparisons state by state. An approach under consideration by the federal government is to transform the measurement practices of the National Assessment of Educational Progress (NAEP) so that state-by-state comparisons may be possible. NAEP has been administering measures periodically to US students in reading comprehension, writing, and mathematics on a regular basis. At the present time, the administration of these measures allows for interpretation by broad geographical region, rather than for each state. The proposal calls for administering these measures so that a

representative sample of each state would be tested and described in NAEP reports. The proposal also expands the number of subject matters assessed. If accepted, this approach could focus the evaluation of schooling on the NAEP achievement measures. Is this a good thing? There is a clear division of opinion. Let me review some of the arguments on behalf of and against this approach. On the positive side:

- A common basis for understanding student achievement would be systematically available.
- The quality of measures would continue to improve because of the salience of the measures.
- States could use such information for their own policy assessment to check their progress.
- Interpretation for policy purposes would be simplified.
- States would be able to compare themselves to subsets of other similar states.

On the negative side, critics contend that:

- NAEP may turn into a national achievement test, and a national curriculum may follow.
- NAEP will not be sufficiently responsive to local or regional differences in curricula, students, or economic factors to permit legitimate comparisons.
- NAEP will drive out state and local tests, which are more responsive to local curricula.
- The pressure for school district comparisons will follow state comparisons.

- Because NAEP's strength will be comparisons over time, the pressure to keep NAEP measures the same will inhibit new goals for the curriculum and new approaches to measurement.

- A single set of measures can be wrong. Given the state of understanding of achievement measurement, investing in different assessment approaches is the most prudent way to collect policy relevant information.

For each of these points, both positive and negative, there are counterarguments, and counter-counter arguments. If the problem were simple, it would already be solved. The attractiveness of a clearly understood, single set of measures for American education is strong, even when the validity of the measures for assessing local and state educational policies is questioned. The state-by-state NAEP approach needs to be understood as an attempt to catch hold of what our schools are doing.

Quality indicators. Another tack is the quality indicators movement (Office of Educational Research and Improvement, 1987). The goal of this effort is to identify and systematically collect information that can give a picture of the overall quality of American education, not simply limited to achievement testing. Work in this area has been conducted by The Rand Corporation, the Center for Policy Research in Education at Rutgers, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA, and by numerous other institutions and scholars. Part of its impetus comes from the realm of economic indicators, where seemingly simple numbers like the Gross National Product, unemployment figures, and the Dow Jones Average efficiently communicate the economic health of the country. The Center for Education Statistics (a division of OERI), under the leadership of Emerson Elliott, has been working on indicators of educational quality. These indicators include figures such as dropout rates, per capita student funding, student-teacher

class ratios, enrollment figures, and the like. Problems encountered with this approach include the vastly different reporting approaches taken for something as understandable as dropout. Different districts and states count dropouts at different intervals, for different ages or grades, use different base rates, track student mobility differently, and so on. Getting everyone to agree on a single reporting approach, even for an "easily understood" concept like dropout, is a Herculean task.

Outcomes like achievement test scores, college admission rates, or dropout figures represent the easy part of indicators. Quality indicators should also take into account input variables and measures of process.

Imagine one wanted a "quality indicator" related to some intermediate process, such as student coursework. In fact, UCLA and The Rand Corporation are collaborating of the development of such indicators. We need to consider how to determine "quality" in a valid and comprehensive way, how to collect such information accurately and comfortably in schools, and how to report such findings so that the effects of educational reform can be tracked. If we (or others) can solve such a problem, educational achievement tests can be relieved of the perhaps excessive burden they carry as measures of the effects of different policies. Making changes, such as adding coursework requirements, strengthening the content of the curriculum in a particular area, or requiring textbooks to exhibit certain content standards, are all hypotheses that policymakers make about what will help schools. Indicators of the extent to which these policies are used is a first step; studying the relationship of the level of their use and resultant levels of student achievement is a second critical link. Yet, the indicator movement must be cautious about identifying a single magic index or number to stand for complex educational processes. As Leigh Burstein of UCLA points out, the context in which such data are reported, understood, and interpreted, is central to the success of this effort (Baker, 1987).

Summary

The search for approaches to assess schools, their teachers, and students, will continue. This discussion has touched lightly on a number of complex issues. Controversies will also continue, and we can be sure that almost any decision will be rethought sometime in the future. Our interest in the research community is to keep a few issues in front of the public and the decisionmakers in this area.

First, we believe that the validity of any measure or indicator should be paramount, whether it is a measure of outcomes, like student achievement, of input, like teacher knowledge, or of processes, like student coursework. These measures should be designed in a way to allow multiple or flexible ways to demonstrate success for different students. These measures should help us to pinpoint and fix weaknesses in policy and practice. Finally, these measures first must serve the interests of students and improve their schools. We must overcome our habit of preparing measures for the convenience of test developers, administrators, legislators, or even teachers. Rather, we need to consider the impact of our approaches to assessing educational effectiveness on our current and future students.

CRESST QUALITY INDICATORS STUDY GROUP

Further Thinking on the Merger of the National Assessment of Educational Progress and the School and Staffing Surveys: Summary and Recommendations from Two Meetings of Statisticians and Researchers

TABLE of CONTENTS

REPORT

APPENDICES

Item I. Sample Letter of Invitation to Participate

Item II. CRESST Statement of Questions and Issues

Item III. List of Participants at Wednesday and Friday Meetings

Item IV. Summary of Main Points from Wednesday Meeting

Item V. Follow-up Letter to Meeting Participants

Item VI. Summary Statements from Meeting Participants and Other Consultants

Item VII. Center for Education Statistics Memoranda on Issues Sent to Meeting Participants

Item VIII. Letter to Participants Regarding Summary and Recommendations from meeting

Further Thinking on the Merger of the National Assessment of Educational Progress and the School and Staffing Surveys: Summary and Recommendations from Two Meetings of Statisticians and Researchers

Leigh Burstein
Pam Aschbacher
Center for Research on Evaluation,
Standards and Student Testing
University of California, Los Angeles

This report summarizes the discussions from two meetings (held at the Center for Education Statistics (CES), Wednesday, November 18 and Friday, November 20, 1987) to advise CES regarding the possible merger of the National Assessment of Educational Progress (NAEP) and the School and Staffing Surveys (SASS). It also incorporates points from written statements provided by selected meeting participants and from other individuals whose advice was sought but were unable to attend. The report begins with a brief description of the background and context of the meetings. A summary of the main points of discussion and recommendations to CES follow. The latter are further illuminated by the written statements from participants (ITEM VI in the attachments).

Background

The meetings on merging NAEP and SASS were organized as an activity of the Quality Indicators Study Group of the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles. The CRESST activity was in response to conflicting advice received by Emerson J. Elliott, Director of CES, and his request for assistance in obtaining further thinking about the possible merger.

As described at the outset of the meeting, the dilemma was as follows. CES has 2 major studies serving complementary purposes, both of which are state-representative. NAEP is a study of student outcomes and is newly state-representative (used to be only national-based sample). SASS is a study of teacher demand and shortage based on state-representative data to be fielded for the first time in 1988 but with the intent to develop a time series on important school characteristics. Currently, the studies have different foci with respect to units of observation and analysis (NAEP focusses on students and their teachers and schools; SASS on teachers and the schools and districts within which they work) and consequently different sampling universes. In 1988 (the first year for SASS), the studies will be fielded in non-overlapping schools.

Reasons offered in support of integrating the two studies include:

1) each may provide contextual data to better interpret the other;

2) the merger represents an opportunity to look at relations between two sets of data;

3) there should be cost savings from reducing number of teachers sampled; and

4) it may be possible to reduce overall respondent burden although the burden may increase on some of those sampled.

On the surface, then, it seemed attractive to merge the two. In fact, in previous meetings, the Advisory Council on Education Statistics (ACES) has recommended that a merger of NAEP and SASS proceed. This ACES recommendation was consistent with the recommendations on linking data collections from the report on alternatives for a national data system on elementary and secondary education prepared by Hall, Jaeger, Kearney and Wiley (December 20, 1985).

Yet other segments of the educational community raised many questions about whether the merger was a good idea based on a variety of technical, substantive, practical, and political grounds. These include

-- management concerns associated with two huge data collection efforts and the need to protect NAEP at all costs;

-- lack of sufficient prior experience with SASS to judge how this survey will be most useful;

-- questions of which background data should be related to student-achievement? how significant would this add-on of questions be? Couldn't this be done in smaller studies?

CES has had many meetings and written several papers about a NAEP/SASS merger (Cf. ITEM VII). Since CES needs to field the study in March 1988, there is need for immediate input. Moreover, at the time of the meeting, CES didn't have any information on the overlap and "strain" projected from simulations of sampling for the three major studies (NELS, NAEP & SASS).

The purpose of the meetings was to bring together persons knowledgeable about educational research, statistical, and policy analytic issues (Cf. ITEM III for a list of meeting participants; other individuals were invited but were unable to attend) that CES's data collections (including NAEP, SASS, Longitudinal Studies) to:

a. Consider the range of issues that CES had already identified and review its available documentation regarding these issues;

b. Augment CES's prior analyses with other evidence that bears on the perceived benefits and costs of the proposed merger;

c. Assess the likely consequences (e.g., for knowledge production, enhancing policy analysis capabilities, improving or degrading the quality of data) of the merger;

d. Recommend options with regard to the decision process on the possible merger and the steps that should be undertaken in advance of a final determination to proceed with the merger.

Participants were provided in advance specific questions and issues that the meeting was intended to consider (ITEM II) and a set of pertinent documents (ITEM VII plus copies of CES Working Paper 2, the Hall et al. report, the synthesis of invited papers from the Elementary and Secondary Data Collection Redesign Project, and the report from the planning conference to consider a merger of NAEP and NELS). Three persons (Richard Jaeger, University of North Carolina, Greensboro; Richard Murnane, Harvard University; Marshall Smith, Stanford University) were asked to provide written input even though they were unable to attend the meetings.

Two 5-1/2 hour meetings were held, with a day in between to accommodate the schedules of the desired participants and to allow time to prepare information from the first day's discussion (ITEM IV) to assist the second day's deliberations. Meeting participants were also asked to provide written summary statements regarding their views on the merger. A follow-up letter was sent to all participants on Wednesday, November 24th, to provide an initial summary of the meetings' main points and a preliminary list of recommendations, and to encourage participants to submit written statements and inform them of next steps. As of December 10th, ten meeting participants (in addition to Jaeger, Murnane, and Smith) had provided such statements (See attached ITEM VI).

Summary of the Issues Discussed

Despite the diversity of perspectives and interests represented in both days' meetings, there was considerable consensus about the basic issues that need to be addressed. These issues were also echoed in the written statements. Display 1 represents an attempt to code the written statements with respect to their consideration of the major issues and support for the recommendations.

Display 1

The main issues addressed were as follows:

Issue 1. What does "merger" mean and how comprehensive (with respect to instrumentation and to samples) should it be?

Discussion: The merger of NAEP and SASS could occur in a variety of ways that vary in the extent of combinations. Merger options discussed include:

- o Complete merger -- joint administration to the same sample of schools on the same cycle (See Richard Jaeger letter)
- o Integrate the two studies in only some states
- o Merge the two infrequently (e.g. every 6 years, which could be accomplished by putting SASS on a 3-yr cycle and leaving NAEP on a 2-yr cycle)
- o Move (or repeat) a small set of items on school policy and teacher characteristics from SASS to NAEP in order to explore certain policy issues, allow NAEP to get better information with its own sample, and allow SASS to keep its own sample and purpose. (Note: this approach would not necessitate including "high-burden" items from SASS in a revised NAEP.)
- o Use part of SASS in NAEP schools
- o Merge the two only at the national level.

A repeated theme of the discussion and written statements is that "considerations of how/whether some elements of the two data collections might be usefully integrated should be examined carefully in the light of specific analytic benefits, respondent burden, data objectives, and periodicity of the data collections before a decision to seek merger" (Linda Darling-Hammond). The primary rationales for the merger proposal were the analytic benefits from adding SASS data about districts, schools, and teachers to NAEP data on schools, teachers, and students and the efficiencies of data collection that might be obtained through using the same sample for NAEP and SASS.

There was general sentiment for moving or readministering some SASS questions as part of NAEP and little or no support for complete merger, at least in the near future. However, how far to proceed needed to be guided by the tradeoffs between the analytical purposes such a merger could serve and the possible consequences in terms of burden and costs of the particular form of merger.

DISPLAY 1. PARTICIPANTS' OPINIONS REGARDING ISSUES IN THE MERGER OF NAEP AND SASS[1]

ISSUES	PARTICIPANTS[2]														
	DB	AB	LDH	EH	MH	RJ	TK	DK	RM	DR	PS	RS	MS	BT	DW
1. Should Merger Occur															
A. Complete Merger in 1990	?[3]	?	N	N	N	N	N	N	N	N	N	N	N	N	N
B. Administer subset of SASS with NAEP		Y?	Y	Y		?	Y		Y			Y	Y	?	Y
C. Further Study regarding 1992/1994 needed	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N?	Y	Y
2. Purpose of Merger															
A. Causal Analysis of school effects		Y	N					N	N	N	N	N	N	?	N?
B. Policy Analytic		Y	Y	Y				Y	Y	Y	?	Y	?	Y	Y
C. Access and Participation			Y	Y				Y	Y		?	Y	Y	Y	Y
3. Topics Requiring further Study															
A. Conceptual Analysis of issues that Merged Data would address			Y	Y		Y	Y	Y	Y	Y	Y	Y		Y	Y
B. Empirical Studies with existing NAEP & SASS data		Y	Y	Y	Y	Y	Y	Y		Y	Y		Y	Y	
C. Respondent Burden (Non-Cooperation)	Y	Y	Y	Y		Y	Y		Y		Y	Y		Y	Y
D. Costs of Merger		Y		Y		Y			Y			Y			Y
E. Value of Design of Common Sample Universe				?		?									Y
F. Incentives for School Participation	Y	Y	Y		Y	Y									

	DB	AB	LDH	EH	MH	RJ	TK	DK	RM	DR	PS	RS	MS	BT	DW
4. Other Options and Issues															
A. Modify the Cycle for SASS			Y		Y		Y		?			Y?		Y	
B. Field Parts of SASS on Different Cycles			Y				Y		?			Y?		Y	

NOTES:

1. The codes are derived from a reading of the written statements provided by the participants. No attempt was made to reach a judgement based solely on comments during the meetings. Full text of statements are attached at Item IV.

2. Participants are denoted by initials and are listed alphabetically. The Participants who provided written statements are: David Bayless (DB), Al Beaton (AB), Linda Darling-Hammond (LDH), Ed Haertel (EH), Morris Hansen (MH), Richard Jaeger (RJ), Tom Kerins (TK), Dan Koretz (DK), Richard Murnane (RM), Paul Sandifer (PS), Ramsay Seldon (RS), Marshall Smith (MS), Brenda Turnbull (BT), David Wiley (DW).

3. The codes are as follows:

Y: Yes (affirmatively responded to the issue)
N: No (Negatively responded to the issue)
?: Statement may address issue but unclear

Blank: Did not mention issue

Issue 2. What analytical purposes should guide any merger decisions?

Discussion: Three sets of analytical purposes might serve to justify the merger:

1. Causal Analysis of Effects of Schools and Schooling

2. Policy Relevant Research Issues

3. Topical Policy Issues

There was considerable agreement among participants that "valid analysis/causal modelling" of school effects cannot be obtained through a merger of NAEP and SASS. The resultant survey would still be a cross-sectional one, and a longitudinal survey such as HS&B and NELS is necessary to contribute to this purpose. Moreover, the merged NAEP/SASS would encourage "invalid but potentially influential studies of school effects that could seriously distort policy" (Dan Koretz; statements from Richard Murnane and Marshall Smith make essentially the same point).

The ability to enhance policy analytic capacity (purposes 2 and 3), on the other hand, received considerable support. Improving the utility of both NAEP and SASS as indicator series was considered a valid and powerful reason for consideration of further integration of the two surveys. But such possible improvements should be attempted only if the integrity of NAEP as an indicator of student achievement trends was not threatened. The notion of using NAEP as a source of student performance data for SASS, for example, was not considered a sufficiently compelling reason to merge at this point, especially since there is as yet no history of the functioning of SASS and its niche in the comprehensive education information system to be maintained by CES.

The most discussed and agreed upon purpose for some degree of merger of NAEP and SASS was to enhance NAEP's usefulness in exploring equity issues. That is, the selective inclusion of SASS questions on teaching and schooling conditions could be used to examine differential student assignment to types of teachers and classes ("access"). It was felt that this analysis would be useful at both the state and national levels for both public and private schools. The following questions were raised about even this purpose, however:

1) Would this information actually be used? Some states evidently have such information already and do not use it.

2) Is it properly a federal task to provide such information on a state-by-state basis?

(Letters from Ed Haertel and Marshall Smith convey the issues on this point.)

Two other (non-causal) analyses were proposed briefly: linking student achievement to staffing variables, and linking

teaching strategies to staffing (Dan Koretz's letter provides a useful example here). The national and state-level patterns in these relations over time were of interest to some participants.

The message that came through loudly was that to date there has not been sufficient conceptual and empirical analyses of the specific analytical purposes to be served by integrating the two data bases. Analyses thus far have focussed largely on operational matters (logistics, respondent burden, costs) while general and overly vague purposes remain the source of the urge to merge. Before proceeding too far down the road of a merger of any consequence (other than the augmentation of 1990 NAEP with a few questions from SASS), further study of the specific analytical issues to be addressed through merger is essential.

Another point latent in the discussion in this area was whether some of the substantive reasons put forward as a basis for the merger might be best served through special studies that parallel and piggyback on either NAEP or SASS. This type of linkage is suggested in the RAND Corporation's report to the NSF on alternatives for the development of a comprehensive information system in mathematics, science and technology education (Shavelson et al., 1987). Currently, certain bridging studies are fielded along with NAEP to address special topics. These are conducted on subsamples as part of the overall study. Many of the ideas that warrant special attention could be fielded in a subset of locales, for instance. One participant (Ed Haertel) suggested that state assessments might be a viable of the outcomes data to augment SASS. In this case, these would be be special studies that would add little actual additional local burden, especially if linkage were carried out at the state level.

There was considerable sentiment from the entire group that CES needs to encourage and commission conceptual and empirical analyses from a broader audience to assist in their development of analytical purposes for integrated data collection. In particular, mechanisms that would encourage empirical studies with already collected NAEP teacher and school data and with the 1988 SASS data are critical if the possibility of a significant merger remains a consideration for the agency.

Issue 3. What are the likely consequences of merger alternatives with respect to respondent burden and costs?

Discussion: The participants expressed a concern that there are so many possibilities and so many assumptions and variables that affect cost, that the cost issues are not clear at all. For example, merging the two studies might imply that all the data in the merged portion should be collected with the same method (i.e. personal interview or mail survey). While it may be expected that the NAEP interview method is less expensive, it might be prohibitive when used with a sample the size of the SASS. The choice among modes of data collection within the merged survey was viewed to be considerably important with respect to both the cost and burden issues (e.g., see statements from David Bayless and Morris Hansen). In particular, the savings from merging SASS is slight given the current plans calling for a mail survey.

A major concern throughout the meeting was the effect of burden (actual and perceived) on quantity and quality of data. It was pointed out that these effects might be exacerbated over time when the data is repeatedly collected every couple of years (see David Bayless and Richard Jaeger statements). Most of the item overlap of the two studies falls on school administrators and teachers. However, district administrators would also perceive increased burden as the number of participating schools or amount of participation by any one school within their district increases. The issue is more one of politics than loss of instructional time.

Since districts differ so, it is expected that they may react differently to the burden of a merger. Some might elect to test a universe of districts since so many may be sampled, whereas others may elect not to participate at all. It is feared that most districts are small enough that the increased burden would discourage them from participating. It was agreed that the 1988 data collection efforts in NAEP and SASS separately will provide some basis for estimating the burden of a partially merged study in the future. Many participants (in particular, David Bayless, Al Beaton, Linda Darling-Hammond, Ed Haertel, Richard Jaeger, Brenda Turnbull) strongly urged more systematic study of respondent burden options before proceeding with anything beyond a mild data linkage.

The group discussed the desirability of providing some payoff to the districts and schools for participating; however, no individualized reports or products seemed appropriate given the sampling methods. In addition, it was pointed out that providing any individualized information to participants in a timely manner would also be difficult. While the question of appropriate incentives was recognized, there was much divergence of opinion on how to respond to this concern. (Note, for instance, statements from David Bayless, Linda Darling-Hammond, Morris Hansen, Richard Jaeger, and Paul Sandifer).

Issue 4. How does the question of the desirable/necessary cycle/periodicity and timing of SASS (or parts of SASS) interact with the above?

Discussion: There was considerable expression of concern that steps be taken to reduce the data collection burden within a given year in some manner. Moreover, these concerns were typically linked with the question of whether it was necessary to collect all, or any part, of SASS on a two-year cycle, especially with the possibility of augmenting NAEP with some SASS questions and sample enhancements (given the differences in target teachers of the two surveys). Participants' statements most clearly articulating the issues here are from Linda Darling-Hammond, Morris Hansen, Tom Kerins, Paul Sandifer, and Brenda Turnbull. While an as yet undetermined "core" data set may be needed every 2 years, much of the SASS data could be collected less frequently. Hansen and Sandifer, in particular, urge that NAEP and SASS be administered in alternative years where a 2-year interval is considered to be essential.

The notion of a 3-year or 4-year cycle for SASS had considerable appeal for a number of participants. Putting SASS on a 3-year cycle would have the advantage of making it coincide with NAEP every 6 years, thus providing possibilities of obtaining some merged data without increasing the burden most years. A 3-year cycle would also allow additional time to analyze the 1988 SASS data before decisions regarding its next administration. Such a choice would also postpone the merger decision to a point beyond the first planned comprehensive state-level data collection.

It was pointed out that collecting some of the SASS data less frequently might provide some funds for collecting other data (e.g. collecting SASS district data every 4 years and collecting finance data on alternate occasions) or for conducting some special studies (See Darling-Hammond, Kerins, and Turnbull statements). Here, again, special studies using existing data bases were considered essential, and some means needs to be found to ensure that they are conducted.

Issue 5. What sets of analytical exercises/special studies should be undertaken to address the merger issue in both the short run and the long run?

Discussion: There are so many unknowns -- changes in NAEP in the future and SASS being completely new -- that it is difficult to think precisely about what would happen if the two were merged in some way. One important step is to use the 1988 fielding of both as a pilot for each separately that will provide some basis for estimating the consequences of any sort of merger. It was agreed that adequate preparation for a merger would entail postponing the merged data collection until at least 1992-1994, especially given the amount of lead time necessary for OMB clearance and so forth. Several additional suggestions were made (virtually all participants had specific suggestions for special studies):

a) CES should do some futures projections to see the costs and consequences

b) There should be some small analyses contracts to look at the NAEP "public useful" tapes regarding the analytical value of merging parts of SASS with NAEP.

c) It would be useful to consider the various augmentations and analyze the incremental value of one over the others.

d) Possible pilots for the merger could involve only one or a small number of states, or perhaps only merge first at the national level.

One proposal for a special study that warrants special mention dealt with the development of a common sampling frame. David Wiley suggested that CES might consider giving up current NAEP and SASS sampling frames and design a new one to integrate both (e.g.,in 1992 or 1994). Then subdivide students and teachers according to whether they are in the NAEP sample universe, and draw separate subsamples and collect some linked data. This idea might be particularly valuable when NAEP becomes state-representative. Other participants thought that a feasibility and cost study of this idea would be worthwhile. However, it would make the most sense when less than the full SASS is fielded. There was a belief that the 1988 experience with the heavy burden in the field without merger might be informative on this matter. The question of mail vs. in-person survey would also impact on this decision.

Major Recommendations

The broad outlines of the recommendations from the meetings are evident from the discussions of the issues and the statements provided by the participants. The recommendations that achieved a general consensus from the meetings and written statements are:

1. A complete merger of the questionnaires and samples from NAEP and SASS should NOT be attempted in 1990. The risks of overburdening NAEP in 1990 are too great; Moreover, too little is known about how SASS will actually function at this time to assess the benefits and consequences of strong ties with NAEP.

The group consensus was that a complete merger (joint administration on same cycle in same sample) is not feasible in 1990 and probably is not a good idea anyway. The purposes (and the samples) of the two studies are legitimately different and should be preserved. Although it might be possible to define a common sampling frame, this approach might be quite inefficient and might have very negative consequences among schools and districts due to its perceived and actual burden. There was interest, however, in the possibility of a partial merger based on the desire to explore the issue of student access to teachers.

2. Whether NAEP and SASS should merge in 1992 or 1994 warrants further study including analyses of existing data from the two surveys gathered through the 1988 data collection.

3. Regardless of the extensiveness of the eventual merger, the analytical purposes that should guide merger efforts should be those dealing with informing policy analysis rather than enhancing capabilities to conduct school effects or effectiveness research in an integrated national or state-representative data base. Examples of policy analytic purposes that could be served through a "merger" effort are the gathering and maintenance of national (and perhaps state representative) indicator series dealing with questions of access and participation (e.g., which kinds of students receive instruction in which kinds of schools from which kinds of teachers?)

4. For the short term (e.g., 1990), a small set of teaching and schooling conditions questions selected from SASS could be administered with NAEP to enhance its ability to serve policy analytic purposes. To this end analytical work using past NAEP collections of teacher and school characteristics as well as other efforts to identify specific policy analytic purposes to be served should be carried out in time to modify and augment the 1990 NAEP school and teacher characteristics questionnaires.

5. A three-year or even a four-year cycle for the major SASS data collection should be considered with at least part of the resource savings shifted to conducting special studies (e.g., longer term study of flow of teachers into and out of the workforce for a panel of schools and districts; augmentation of NAEP data collection in 1990; studies of the consequences of the intensity of respondent burden and costs consequences of major merger). Alternatively, the SASS instrumentation can be broken up into smaller sets which could be fielded on different cycles with perhaps a core set maintained on a more frequent cycle. Spreading out the SASS cycle would also postpone collection activities in ways that would place less strain on plans for the

1990 NAEP.

6. Postponing major merger discussions beyond 1990 provides time and resources to consider (through design and other special studies) the costs and benefits of developing a merged sampling universe across the major data collections (including NELS as well as NAEP and SAS).

7. Attention is needed to the benefits accrued at the school level from participating in these surveys. "Contributing to national well-being" is an increasingly weak incentive given the extensiveness of data collection demands and competition from data collection with greater extrinsic rewards.

* * * * *

The above summary and recommendations convey the tenor of the discussions and written statements. Participants were genuinely concerned that the primary purposes of NAEP and SASS not be sacrificed or damaged by a hurried decision to merge the two. CES is undertaking major modifications and extensions of its data collection responsibilities over the next few years. Its efforts to date are commendable and the general direction of agency was viewed positively by the participants. Nevertheless, under the circumstances of major changes in responsibilities, operations, resources, and staffing, time and resources devoted to further study that enhances the likelihood of fielding and reporting these collection efforts in an effective and credible manner is critical. Discussions of mergers of these data collections need to proceed at a more deliberative pace than at present. There is just too much at stake.

December 10, 1987

CRESST QUALITY INDICATORS STUDY GROUP

Report from Meetings on CES Merger of NAEP and SASS

ITEM I

Sample Letter of Invitation to Participate

UNIVERSITY OF CALIFORNIA, LOS ANGELES UCLA

BERKELEY · DAVIS · IRVINE · LOS ANGELES · RIVERSIDE · SAN DIEGO · SAN FRANCISCO SANTA BARBARA · SANTA CRUZ

CENTER FOR THE STUDY OF EVALUATION
CENTER FOR RESEARCH ON EVALUATION, STANDARDS AND STUDENT TESTING
UCLA GRADUATE SCHOOL OF EDUCATION
405 HILGARD AVENUE
LOS ANGELES, CALIFORNIA 90024-1521
(213) 825-4711
(213) 206-1532

November 6, 1987

Dr. David Wiley
227 Sheridan Road
Kenilworth, IL 60043

Dear David:

Thank you for agreeing to participate in the examination of technical issues in the possible merger of the National Assessment of Educational Progress (NAEP) with the School And Staffing Survey (SASS). This examination is being conducted as an activity of the Center for Research on Evaluation, Standards, and Student Assessment's (CRESST) Study Group on Quality Indicators to assist the Center for Education Statistics (CES) in its deliberations of the merger question. The CES Advisory Council has recommended that the merger proceed. Other segments of the educational community have questioned the advisability of the merger on a variety of technical, substantive, practical, and political grounds.

The purpose of this examination is to:

a. Consider the range of issues that CES has already identified and review their available documentation regarding these issues;

b. Augment CES's prior analyses with other evidence that bears on the perceived benefits and costs of the proposed meger;

c. Assess the likely consequences (e.g., for knowledge production, enhancing policy analysis capabilities, improving or degrading the quality of data) of the merger;

d. Recommend options with regard to the decision process on the possible merger and the steps that should be undertaken in advance of a final determination to proceed with the merger.

The plan of operation for the present activity is to seek advice in two ways with respect to these issues and a specific set of questions regarding the merger (see enclosed). The primary means will be through two meetings to be held at the Center for Education Statistics in Washington on Wednesday, November 18, 1987, and Friday, November 20, 1987. In addition

written reactions will be obtained from a select set of individuals unable to attend either meeting (their written input will be due by November 30, 1987). Participants will include researchers and policy analysts knowledgeable about the examination of educational effects through large-scale data analysis, experts in survey sample design, and representatives from national, state, and local organizations with an interest in analyses of education and the conduct of major survey data collections in the schools. The discussions at the meetings and the written reactions will be synthesized into a set of recommendations to CES about viable next steps and their possible consequences.

We are able to offer a modest honorarium and travel expenses for participating in the scheduled meetings. Included for your completion and signature is a Consultant Agreement. Please return the signed agreement in the enclosed self-addressed along with your current vita.

We will be contacting you shortly to assist in travel arrangements and local hotel accommodations and to notify you about the exact schedule and location for the meetings.

A set of papers and reports that serve as background reading for the discussions is enclosed. At this point the short issues papers and Working Paper #2 perhaps represent the most pertinent if your reading time is restricted.

Thank you in advance for your willingness to participate.

Sincerely,

Leigh Burstein
Co-Director, CRESST
Quality Indicators Study Group

Eva L. Baker
Co-Director, CRESST
and Co-Direct CRESST
Quality Indicators Study Group

CRESST QUALITY INDICATORS STUDY GROUP

Report from Meetings on CES Merger of NAEP and SASS

ITEM II

CRESST Statement of Questions and Issues

UCLA
CENTER FOR THE STUDY OF EVALUATION
152104

11/4/87

CRESST Quality Indicators Study Group

Meetings on CES Merger of NAEP and SASS
November 18 & 20, 1987

Questions and Issues

1. What analytical advances are afforded from the combination of the samples for NAEP and SASS?
   a. with respect to enhancing the analysis of schooling? instruction? teachers and teaching?

   b. Does the existence of a national sample consequentially enhance the analyses identified in a.?

   c. Would the existence of data on a state-by-state basis consequentially enhance the analyses identified in a.?

   d. Do the presumed advances represent a unique opportunity or simply augment existing efforts (e.g.?) in a significant way?

   e. What are the consequences for other data collections designed to address related issues?

2. Is it possible to merge the two national samples without adversely affecting the quality of the data to address the primary questions the data sets were designed to examine?

   a. Will the resultant respondent burden compromise the quality of data for assessing educational outcomes from NAEP and schooling conditions from SASS?

   b. Will the compromises in sample selection and design consequentially impact each of the separate collection efforts?

3. If the decision were made to proceed with the combination, how would one carry it off given the distinctions in the primary purposes and sampling between the two studies?

   a. What should be the stages for phasing in the combination, keeping in mind the planned expansion of NAEP in 1990?

   b. What set of special studies, pilot studies, and simulations should be carried out before a final decision to proceed with the merger (re. pilot test 1989)?

   d. What is a reasonable timeframe in light of data collection cycles for conducting studies of the merger before a final decision is made?

Item III December 10, 1987

CRESST QUALITY INDICATORS STUDY GROUP

Report from Meetings on CES Merger of NAEP and SASS

ITEM III

List of Participants at Wednesday and Friday Meetings

CRESST Quality Indicators Study Group

Meetings on CES Merger of NAEP and SASS

Participant List

Wednesday, November 18, 1987

Non-CES

Pam Aschbacher, CRESST, UCLA
Eva Baker, CRESST, UCLA
Anthony Bryk, University of Chicago
Leigh Burstein, CRESST, UCLA
Joe Conaty, OERI
Morris Hansen, WESTAT
Dan Koretz, Rand Corporation
James McPartland, CSOS, Johns Hopkins
Senta Raizen, National Research Council, National Academy of Sciences
Paul Sandifer, South Carolina State Department of Education
William Schmidt,Office of Studies and Program Assessment, National Science Foundation
Ramsay Selden, CCSSO State Education Assessment Center
Brenda Turnbull, Policy Studies Associates

CES Staff

Emerson Elliott
Jeanne Griffith
Anne Hafner
Carrol Kindel
Don Malek
Eugene Owen
Mary Papageorgiou
Gary Phillips
Paul Planchon
Iris Silverman
Nancy-Jane Stubbs
David Sweet
Doug Wright

CRESST Quality Indicators Study Group

Meetings on CES Merger of NAEP and SASS

Participant List

Friday, November 20, 1987

Non-CES

Pam Aschbacher, CRESST, UCLA
David Bayless, WESTAT
Al Beaton, NAEP, ETS
Leigh Burstein, CRESST, UCLA
Joe Conaty, OERI
Linda Darling-Hammond,RAND Corporation
Ed Haertel, Stanford University
George E. Hall, Slater Hall Information Products
Tom Kerins, Illinois State Department of Education
Sally Kilgore, OERI
Doris Redfield, OERI
Ramsay Selden, CCSSO State Education Assessment Center
W. Ray Turner, Dade County Schools, Miami, Florida
David Wiley, Northwestern University

CES Staff

Emerson Elliott
Jeanne Griffith
Carrol Kindel
Don Malek
Eugene Owen
Mary Papageorgiou
Gary Phillips
Paul Planchon
Iris Silverman
Nancy-Jane Stubbs
David Sweet
Doug Wright

Participants Providing Written Input Only

Richard Jaeger, University of North Carolina, Greensboro
Richard Murnane, Harvard University
Marshall Smith, Stanford University

CRESST QUALITY INDICATORS STUDY GROUP

Report from Meetings on CES Merger of NAEP and SASS

ITEM IV

Summary of Main Points in Wednesday Meeting

CRESST Quality Indicators Study Group

Meetings on CES Merger of NAEP and SASS
November 18 & 20, 1987

MAIN POINTS IN DISCUSSIONS
WEDNESDAY, NOVEMBER 20, 1987

The group attending the meeting on Wednesday considered the original list of questions and issues that were distributed in advance of the meeting. The main points raised in those discussions included the following:

1. What is meant by a "merger" of NAEP and SASS is subject to a variety of interpretations. Strong merger implies joint administration on a repeating cycle in the same sets of schools. Weak merger can be accomplished in a lot of ways with the most benign and obvious being move toward comparable wording where current intents overlap and inclusion of additional SASS-type questions within the NAEP schooling conditions data collection.

2. There was a strong commitment that the primary purposes of NAEP and SASS should be preserved at all costs. Any risks to those purposes should be avoided. Preserving the outcome series from NAEP nationally and establishing the teacher characteristics and flow series (on a state basis) were considered to be of greatest importance.

3. A strong merger of NAEP and SASS for the primary purpose of improving the relational analysis of the impact of schooling conditions on student outcomes would be a mistake. Basing relational analysis of the causal effects of schooling conditions on cross-sectional studies is a bad idea (misleading is the mild form of the criticism; longitudinal studies are essential for such analyses.

4. Improving and modifying NAEP data collection in the schooling domain to provide better "descriptive" analyses of trends is potentially of benefit as is the possibility of presenting evidence on the relation of student characteristics to the characteristics of the school conditions they receive. But more preliminary investigation is needed to determine just what types of enhancements in the descriptive capacity of NAEP are worthwhile. Moreover, while there may be some justification for national samples for such purposes, the additional benefits of state-level samples for these purposes are less clear. Support for this point implies enhancing NAEP's data collection without moving toward major merger.

5. There was much sentiment for modification of the "perceived" plans for the administration cycle for SASS rather than pushing toward strong merger. The primary argument was the plans (and presumed strong merger) would force more frequent fielding of SASS than is viewed to be necessary for its primary purpose. Expanding the period between administration of SASS was

strongly recommended by some participants. Administering SASS out of phase with NAEP was also proposed on the grounds of the potential respondent burden. While the 1988 fielding of SASS is firm, there was some support for going at least 3 years before repeating this collection. Besides the concern for burden, there was a strong interest in fully developing what is a new initiative without complicating both it (SASS) and NAEP (assuming state level data collection in 1990).

6. More attention should be paid to planning the kinds of special studies that would inform decisions down the road about data linkages than to the push for 1990 merger of the main CES collections. Such studies should include investigations of the respondent burden from more intensive collection within the cross-sectional surveys (implicit in the NAEP-SASS strong merger).

7. More attention should be paid to the questions of benefits to participating districts, schools, and teachers. Arguments of intrinsic merits of serving national interests are insufficient in light of competing data collection burdens.

8. The question of partial paneling of SASS and perhaps NAEP needs further exploration.

CRESST QUALITY INDICATORS STUDY GROUP

Report from Meetings on CES Merger of NAEP and SASS

ITEM V

Follow-up Letter to Meeting Participants

UNIVERSITY OF CALIFORNIA, LOS ANGELES UCLA

BERKELEY · DAVIS · IRVINE · LOS ANGELES · RIVERSIDE · SAN DIEGO · SAN FRANCISCO · SANTA BARBARA · SANTA CRUZ

CENTER FOR THE STUDY OF EVALUATION
CENTER FOR RESEARCH ON EVALUATION, STANDARDS AND STUDENT TESTING
UCLA GRADUATE SCHOOL OF EDUCATION
405 HILGARD AVENUE
LOS ANGELES, CALIFORNIA 90024-1521
(213) 825-4711
(213) 206-1532

November 24, 1987

TO: Participants, CRESST NAEP-SASS Merger Discussions

FROM: Leigh Burstein

RE: Next Steps

Now that I am back, I want to take the opportunity to thank you for your participation thus far in the CRESST-sponsored discussions on merging NAEP and SASS. The sense I have gotten from both CES and some of you is that the meetings went very well. The major issues were aired and received thorough, if not always extensive, discussions. I feel confident that we will be able to provide CES with a set of recommendations that can assist their decision process.

As was discussed at the end of both meetings, this timeframe for input from this activity is very short. It was agreed that I would make a presentation to the CES Advisory Council on Monday, December 14th. To this end, I urge each of you who attended either meeting (and those who did not as well) to provide me a brief (2-3 page) statement regarding your summary recommendations by Thursday December 3rd at the latest. This statement could address the issues and questions as raised initially, various points that came up during the discussions, or ideas you had reflecting on the discussions. The summary of Wednesday's main points that was distributed on Friday and Dick Murnane's letter are enclosed to assist you in this next phase.

I thought it would help if I also provide a brief summary of what I thought occurred during the meetings. There was consistency in the issues discussed during the two days; my quick, rough list is as follows:

1. What does "merger" mean and how comprehensive (with respect to instrumentation and to samples) should it be?

2. What analytical purposes should guide any merger decisions?

3. What are the likely consequences of alternatives with respect to respondent burden and costs?

4. How does the question of the desirable/necessary cycle/periodicity and timing of SASS (or parts of SASS) interact with the above?

5. What sets of analytical exercises/special studies should be undertaken to address the merger issue in both the short run and the long run?

My sense was that while the emphases on the two days differed considerably, there was a general consensus that

1. A major merger of the questionnaires and samples from NAEP and SASS should not be attempted in 1990. Whether such a merger should occur in 1992 or 1994 warrants further study including some basic analyses of existing data from the two surveys gathered through the 1988 data collection.

2. Regardless of the extensiveness of the eventual merger, the analytical purposes that should guide the decision process should be those dealing with informing the policy analytic process rather than the enhancement of capabilities to conduct school effects or effectiveness research in an integrated national or state-representative data base. Examples of policy analytic purposes that should be supported through any "merger" effort are the gathering and maintenance of national (and perhaps state representative) indicator series dealing with questions of access and participation (e.g., which kinds of students receive instruction in which kinds of schools from which kinds of teachers?)

3. In the short term, careful consideration should be given to drawing from the SASS instrumentation teaching and schooling characteristics and conditions questions that would enhance NAEP's ability to serve policy analytic purposes. To this end analytical work using past NAEP collections of teacher and school characteristics as well as other efforts to identify specific policy analytic purposes to be served should be carried out in time to modify and augment the 1990 NAEP school and teacher characteristics questionnaires.

4. Certain functions of SASS do not require two-year cycles. A three-year or even a four-year cycle for the major data gathering of SASS should be considered with at least part of the resource savings shifted to enhancing certain special studies (e.g., longer term study of flow of teachers into and out of the workforce for a panel of schools and districts; augmentation of NAEP data collection in 1990; studies of the consequences of the intensity of respondent burden and costs consequences of major merger).

5. Postponing major merger discussions beyond 1990 provides time and resources to consider (through design and other special studies) the costs and benefits of developing a merged sampling universe across the major data collections (including NELS as well as NAEP and SAS).

There were other points that might appear in the summary recommendations I will prepare for the CES Advisory Council and circulate among participants. We will also prepare a longer report from the meetings.

My plan is to draft the summary recommendations for the CES Advisory Council and circulate them to you (along with copies of the written statements from participants) by December 10th. Any suggested changes will need to be offered immediately to impact the version to be presented to the CES Advisory Council on December 14th. I also expect to attach the written statements as appendices to the summary recommendations unless there is objection.

That's about it for now. If you have any thoughts on the above and would like to discuss them with me, please call. I will be in town through December 6th (213-825-1889; 818-883-9185). Thanks again for your participation.

CRESST QUALITY INDICATORS STUDY GROUP

Report from Meetings on CES Merger of NAEP and SASS

ITEM VI

Summary Statements from Meeting Participants and Other Consultants

1.	David Bayless	Weststat
2.	Albert E. Beaton	NAEP, ETS
3.	Linda Darling-Hammond	The RAND Corporation
4.	Edward Haertel	Stanford University
5.	Morris Hansen	Westat
6.	Richard Jaeger	University of North Carolina
7.	Thomas Kerins	Illinois Board of Education
8.	Daniel Koretz	The RAND Corporation
9.	Richard Murnane	Harvard University
10.	Doris Redfield	OERI
11.	Paul Sandifer	South Carolina Dept of Education
12.	Marshall Smith	Stanford University
13.	Brenda Turnbull	Policy Studies Associates

Merger of NAEP and SASS
The Relationship of Risks of Non-Cooperation
to Level of Commitment and
Total Level of Data Collection Burden

by:

David L. Bayless
Westat, Inc.

A major obstacle to any large-scale national study is the projected lack of cooperation of the sample members among the various levels of the educational system if the data collection activities of the study become too burdensome or there is a lack of commitment. The educator's primary purpose is to manage and deliver instructional services which is in natural conflict with or an obstacle to providing assistance in the collection of data for a research study. Political, philosophical and other factors also contribute to the lack of cooperation.

Several factors (causes) are related to, or associated with, the educator's decision to participate in the data collection activities of the study. It is hypothesized that the risk of non-cooperation is related to: (1) the level of commitment to the study felt by the educators and/or group of educators (e.g., CCSSO-CEIS), and (2) the total level of data collection burden of the current study(s) being implemented. The total level of data collection burden is measured by the length of time for the respondent to respond to the study instruments and the operational time to collect the study data which include the sample size plus the total level of data collection burden of other data collection activities the education unit (state, district, school) has or is commited to collect (e.g., statewide and local achievement testing programs, other university and National, state, and local research studies). Although considerable work is needed to understand the theoretical relationship between the risk of non-cooperation and the level of commitment and the level of data collection burden, based upon my practical experience in several nationwide data collection activities, I have concluded this relationship is important to understand before considering whether NAEP and SASS should be merged.

CES should fund research in this area by gathering data on this relationship so that emperical findings can guide the planning of mergers of large national studies such as NAEP and SASS. In studying this relationship, other extraneous or blocking factors such as sectors of the educational system (public vs private) and level of the educational system [National, state, district, elementary or secondary (schools, teachers, and students)] are factors that should be incorporated into the model. Modern methods of statistical design, such as the design of experiments, should be considered in the study of the relationship between risk of non-cooperation and commitment and burden.

In the absence of valid and reliable instruments to measure and collect data about the relationship of the risk of non-cooperation to the level of commitment and data collection burden, I offer the following comments and observations that CES should take account of considerations concerning the potential merger of NAEP and SASS.

From the beginning of NAEP (1969) to the present day, the priority for the NAEP sample design is to produce National estimates (not state-by-state estimates) for the nation and specified sub-populations and reporting groups. Cooperation with the NAEP survey has been voluntary at the state, district, and school levels. Natural conflicts between the data collection burden of NAEP with the burden of other National, state, and local data collection activities has

existed and will continue to exist. The level of commitment by educational executives to NAEP has been adequate primarily because the data collection burden has not been excessive. Under the proposed sampling plan in 1988, 44 percent of the states will have over 50 percent of their districts in at least one of the National data collections of NAEP, SASS, and NELS. It is my prediction this increased data collection burden will raise the risk of non-cooperation and will affect the quality of the collected data. Let me illustrate this view in relation to the data collection activities of SASS.

The data collection method for SASS of 1988 is to be conducted via a mail survey, which will add an extra data collection burden to the schools, (e.g., school personnel will expend time to "see to it" that the data are collected), which in most cases is a cost to the local school system. Also the data collection burden is at a very high level in terms of the number of sample members selected. Concern has to be expressed about cooperation or the risk of non-cooperation beyond 1988 in important National studies such as NAEP and NELS where much of the operational burden to collect the data is conducted by a person external to the school. Damage could be done to quality of the data for these other studies in future years.

If a priority is to maintain the National data collection activities of NAEP and/or SASS as "pure", i.e., a high cooperation rate and data that conform to strict statistical data collection standards (quality), then only those states whose level of commitment is high should be invited to "piggyback" onto the NAEP and/or SASS sample. States whose level of commitment is low and/or whose total data collection burden (e.g., large state and/or local assessment programs) is large should not be a part of the state level NAEP or SASS studies. Such a plan would reduce the natural conflict that exists between the National data collection activities and the state and district data collection activities and improves upon the "volunteerness" of the data collections tasks to the educational unit at the state, district, and school level.

If state-by-state estimates are required for NAEP and/or SASS, then, in my view, to maintain high appropriate cooperation rates and data quality the data collection activities need to become a legal mandate as they are in certain state assessments (e.g., study participation is not voluntary). If this is to be the case, then I would strongly recommend that concerned state and local school officials be an integral part of making the data collection activities a legal requirement. CES should research this issue by assessing the preferences and opinions of the state, district and school officials (both private and public) in 1988 as to the practical concerns about a legally mandated data collection at a level that will provide separate data by states.

---

* See Table 4 of CES September 17, 1987 tabulation

On Merging NAEP and SASS

Albert E. Beaton
November 25, 1987

In my view, merging the National Assessment of Educational Progress (NAEP) and the School and Staffing Survey (SASS) would somewhat enhance both data bases by including data derived from one in the data base of the other. However, the details of such a merger are very important because a rush to merge might result in decisions which destroy the integrity of either or both of the individual surveys.

There are useful purposes for the merged data. For example, it would be useful to have more detailed information about the teachers of highly performing science students and about the teachers in schools where the average science performance is high. It would also be useful to know more about the teachers of students who have highly or poorly educated parents. Information about teacher and student attributes is useful in describing how resources are allocated. We expect future NAEP reports to include some teacher information, but more would be better. Granted, those who analyze such data will have to be careful not to attribute causation to relationships found in survey data, but findings derived from exploring survey data may lead to hypotheses which can be tested by appropriate experiments.

So far, the NAEP teacher data have been used by Longford, Johnson, and King to explore the question of the amount of student variance associated with teachers and schools using a multi-level model. The results will be presented soon. Multi-level models, such as the one proposed by Aitkin and Longford, which was used in this study, and others, show substantial promise for exploring the relationships between teacher characteristics and student performance.

I expect much more use of the present NAEP teacher data in the future. To encourage this use, and other uses, a sample of NAEP data has been placed on a floppy disk and a Primer is being prepared to help secondary analysts use the NAEP data on the floppy disk as well as the full NAEP data base. This Primer shows in detail how to merge and use the student and teacher data; in fact, the first recipient of the NAEP floppy disk appended the teacher data for his students to explore. My belief is that NAEP would be seriously hurt if no teacher data were available. The extended SASS information would somewhat enhance the NAEP teacher data which are already available.

The details of a merger of NAEP and SASS may be difficult to work out, although I do not see any problem that definitely could not be overcome with proper planning and experimentation.

First, it is clear that NAEP and SASS must be coordinated as long as the SASS occurs in the same years as NAEP. Although the two surveys might be able to use mostly different schools (as in 1988), they cannot avoid using many of the same school districts. The spectre of two different organizations requesting cooperation from the same school districts in an uncoordinated way would almost certainly lead to refusals to cooperate and thus the diminution of both surveys.

The question, then, is whether to minimize or maximize the overlap of the samples. Minimizing the overlap spreads a lighter burden over more schools; maximizing places a heavier burden on fewer schools. Merging the two data bases implies maximizing the overlap so that as much information as possible would be in the merged data base. However, we have no way of knowing at this time whether the added burden on the selected schools and teachers would affect the cooperation rates.

Schools are already wary of the intrusion of NAEP. During the 1986 asssessment, we experienced more difficulty than in any time in NAEP's history in gaining the cooperation of the schools to participate. More and more schools are feeling the burden of a variety of testing and research programs and becoming dissatisfied. NAEP is having to exert tremendous pressure and commit to expensive services in order to maintain our traditional response rates.

We should also note some differences between the NAEP and SASS teacher samples. NAEP samples fourth, eighth, and twelfth grade students and 9-, 13-, and 17-year-olds. SASS samples teachers at all grades and thus teachers in schools where NAEP does not; NAEP selects some teachers of its randomly selected students. SASS is intended to make statements like, "11% of fourth grade teachers..." whereas NAEP is intended to make statements like, "11% of fourth graders have teachers who..." While these differences in sample properties can presumably be worked out, it may be that overlapping teacher samples would be drawn for the surveys. The details of the sampling must be satisfactorily worked out before a merger can responsibly proceed.

Another factor affecting a merger is that the details of the 1990 NAEP have not yet been determined, and the pilot study of state-by-state assessment for that year has not yet been funded. And, of course, the details of full state-by-state assessment in 1992, if funded, have not been planned or approved. Present thinking is that NAEP will assess, state-by-state, twelfth grade students in mathematics in 1990, if funded. In 1992, NAEP hopes to assess, state-by-state, all subjects in each of three age/grade combinations, if funding is available.

If state-by-state funding is not available for NAEP, the overlap of samples will probably be slight, and it is doubtful that merging the surveys will have any benefit for either.

On the other hand, when NAEP is fully funded for state-by-state assessments at all of its age/grade levels, coordination in field operations is clearly necessary and the opportunity to merge data bases may be beneficial to both surveys. Only the question of the costs and benefits of merging would remain, and whether those costs will be greater than the benefits.

Therefore, it seems reasonable to attempt to estimate the costs of merging. To do this, I propose a trial run to study the logistics of the merger procedures. Instead of deciding to merge or not to merge, attempt merging in one or a few states in 1990, if possible, or in 1992. Such a trial run would entail working through the details of coordinating the sampling, field work, and data merging. More importantly, the trial run would give an opportunity to study the reactions of decision-makers in the states, districts, and schools to the merged studies and to learn from them how to attain the required data with minimum disruption of the school system. Measuring differences in cooperation rates would be particularly important. After some practical experience, an extensive bridge study might be in order to assure that the continuity of NAEP is not lost.

RAND

December 4, 1987

Linda Darling-Hammond

Dr. Leigh Burstein
Graduate School of Education
University of California
138D Moore Hall
405 Hilgard Avenue
Los Angeles, CA 90024-1521

Dear Leigh:

In response to your request for further comments on the proposed merger of NAEP and SASS, here are mine. In a nutshell, I believe there are several factors that argue strongly against a full-fledged merger of NAEP and SASS, and that make the consideration of any merger in 1990 inadvisable. Considerations of how/whether some elements of the two data collections might be usefully integrated should be examined carefully in the light of specific analytic benefits, respondent burden, data objectives, and periodicity of the data collections before a decision to seek merger after 1990 is made. Per your request, I am also including a very brief discussion of the components of SASS.

There are two distinct and separable rationales for the proposal to merge NAEP and SASS: (1) analytic benefits to be obtained by adding data about districts, schools, and teachers collected in SASS to data about schools, teachers, and students collected in NAEP; (2) efficiency of data collection that might be obtained by using the same samples for NAEP and SASS. A third consideration is the practical feasibility of merging a data collection that is in the process of substantial change (NAEP) with one that is as yet untested (SASS). These are considered below.

Analytic Benefits

I do not see major analytic benefits to be derived from merging the NAEP and SASS samples and instrumentation wholesale. First, much of the data collected in SASS is designed to support analyses of teacher supply and demand and to provide estimates of school and teacher characteristics for the overall population of schools and teachers. In a cross-sectional sample, these data will not prove highly useful for modelling school effects on student outcomes.

Second, those characteristics of schools and teachers that provide

202 · 296 · 5000

2100 M Street NW
Washington, DC 20037 1270

Main Offices 213 · 393 · 0411
1700 Main Street, PO Box 2138
Santa Monica, CA 90406 2138

descriptions of the experiences of students tested by NAEP are largely already included in the NAEP background data collected from school principals and from the teachers who teach the tested students. [Note that the different goals of the data collections require entirely different sampling of teachers. Whereas SASS seeks to describe the population of all teachers, NAEP seeks to describe the characteristics and practices of teachers who teach students assessed in a given year, e.g. the English teachers of those students assessed in the Reading and Writing Assessment in a given year. Thus, the teacher samples cannot be meaningfully "merged".] Where there are particular gaps in the NAEP background data (e.g. insufficient information about the qualifications of teachers), some modification of NAEP instruments would be sufficient to allow analyses of say, the qualifications of teachers who serve students of different types.

Efficiency of Merged Samples

Given proposals to expand NAEP to state sampling and plans to do so in SASS, there is the obvious question as to whether merging the samples would provide less overall respondent burden for the data collections and result in lower costs for data collection. There are three questions here that need to be evaluated:

1. Will concentrating respondent burden on fewer total districts and schools, reduce overall burden? Will it reduce respondent participation or response rates? Reduction of overall burden would require streamlining the data collection instruments for the two studies. Given the relatively low degree of overlap between them, this would I believe result in very little reduction of overall burden, unless some data elements and survey goals are eliminated from one or the other study. This will require hard choices about objectives for either NAEP or SASS that can be given up. Concentrating respondent burden could lead to lower participation rates, as Joe Turner of Dade County suggested at our meeting. Given the increasing reluctance of states and districts to cooperate in federal data collection efforts, this should be an important concern given careful examination.

2. Will merging samples save administrative costs? This is an empirical question about which I believe there is little consensus at the moment. Contractor costs for contacting districts and schools would obviously decrease if the same contractor administered both collections in an overall smaller sample of districts than would be obtained in independent administration. On the other hand, the costs of securing cooperation for a much larger scale activity and managing the complexities of drawing separate samples of teachers (and perhaps in some cases, schools and districts as well, to satisfy the different analytic goals and estimation objectives of the two collections which produce different sampling considerations) will offset the above savings to some unknown extent.

3. To what extent will the analytic goals of SASS and NAEP be met

with the same sampling specifications? As mentioned above, SASS requires representative samples of districts, schools, and teachers to produce estimates of their characteristics and practices overall and for certain specified strata (e.g. districts by size, urbanicity, etc.; schools by type, level, sector, size; teachers by field, sector, level). NAEP requires representative samples of students, usually selected to be highly clustered in a much smaller sample of schools (since estimates of school characteristics are not the major focus of the data collection), with oversampling of schools by ethnicity and other characteristics of students served in order to support estimates of student achievement for particular subpopulations. Though it is not technically impossible to design samples that serve both goals or to weight the resulting data to serve the purposes of different analyses, the trade-offs or inefficiencies in sampling require examination before the cost savings of merged samples can be assessed.

Practical Feasibility

A major consideration in the decision as to whether some merger is desirable is what the nature of these two data collections will be in 1990. Proposals to revise NAEP, currently being considered in Congress, include expansion to state sampling and possible local add-ons, changes in both the nature and frequency of assessment in various subject areas, and changes in the governance structure of NAEP. Other proposals that have been raised by the Alexander-James Commission and the National Academy of Education may be further pursued by the new governing body of NAEP. These include making NAEP a longitudinal rather than cross-sectional assessment, expanding the (undefined) policy analytic capacity of NAEP, extending its capacity to support analyses of school effects, changing the scaling and reporting features of the assessments, and others. Over the next few years, substantial changes will be made to the design and conduct of NAEP which will totally alter the nature of the data collection activities and will reframe the questions about the desirability or feasibility of merger with any other data collection system. Plans to merge NAEP with SASS will be shooting at a moving target.

At the same time, the first fielding of SASS in 1988 will produce substantial information about changes required in the management of that equally mammoth data collection activity. However, analyses of the initial experiences with SASS will not be available until at least 1989, past the point when planning for a 1990 merger would have to have been well underway. Indeed, a very important goal for the Center is establishing the periodicity of major data collections in such a way that past efforts can inform the subsequent data collections, that time for adequate field testing and analysis of field test results is permitted, and that energies can be devoted to data analysis as well as data collection. Finally, SASS has a number of different components which, though currently joined, may not need to be maintained in tandem in future data collections. Thus, many different options are available for achieving data collection goals short of either full merger, on the

one hand, or simultaneous independent fieldings of NAEP and SASS every two years, on the other.

Components of SASS

SASS currently includes surveys of school district administrators, and public and private school principals and teachers in linked samples of districts and schools. A follow-up survey of teachers in the year after the baseline survey is also planned to track teacher mobility and attrition and to compare leavers to stayers. The data set is designed to support analyses of teacher supply and demand (data elements for these analyses are lodged in each of the district, school, and teacher surveys); and to describe school programs and services, teacher and administrator characteristics and working conditions, and school staffing patterns for different states, sectors, and levels of schooling.

Fielding the surveys with all of these respondent groups and data elements joined is a useful strategy in the first year of implementation (1988) because it permits continuing time-series for some data elements (e.g. counts of teachers by field, and teacher demand and shortage estimates) while launching some new time-series for data that are much-needed but have not been collected by CES in the recent past (e.g. estimates of teacher turnover, characteristics of the teaching force). In addition, some multi-level analyses are made possible by the linked samples of districts, schools, and teachers. However, the surveys may not need to be conducted in precisely the same form or packaged precisely in this way each time.

There are many possibilities for decoupling elements of SASS depending on how often certain kinds of data are needed and whether all of the data elements are necessary for state-level analyses on a regular basis. For example, the Center has already considered using the district survey to collect data on teacher demand and shortage on an alternating basis with data on district finance and expenditures. Data on teacher attrition rates, mobility, and sources of supply can be collected from a few items in the school survey if they are needed on a more frequent basis than other data elements. (Given the burden and costs associated with the full SASS data collection, RAND had recommended this strategy as an option in designing the survey.) State estimates may not be needed for every state in each cycle; samples could be drawn to provide national and regional estimates regularly and state estimates for a rotating third (or some other fraction) of the states during each data collection. Data on school programs and services may not be needed with the same periodicity as data on teacher characteristics. And so on.

In my opinion, a full fielding of SASS on an every two-year cycle is probably not needed and may push the limits of the Center's capacity. Such a cycle allows almost no time for refinement of the survey design based on analysis of the prior cycle's data and data collection experience, and virtually eliminates the possibility of field testing

any proposed changes. Given that most of the information provided by SASS has not been collected or reported for many decades, a 3-year cycle may prove sufficiently timely. Alternatively, staggering the number of states for which representative samples will be drawn or the data elements that will be included in each 2-year cycle could also reduce overall costs and respondent burden. The point is that when considering the costs and benefits of data collection strategies or the possibilities for merging some aspects of different data collections, it is useful to consider a variety of options for meeting various data collection and reporting goals, rather than thinking of SASS (or NAEP for that matter) as a single giant blob.

A Note on NAEP

I believe that some of the rationale for merger on analytic grounds derives from lack of familiarity and use of the full NAEP data set, including its school and teacher survey components. There is also a fair amount of variability in the content of the data set from one assessment to the next. Each panel has its own views on what is important to measure. Given the changes in each assessment in the nature of the teacher samples and the types of background questions asked of school staff as well as students (and the changes in item sampling strategies that have influenced what kinds of analyses can be performed), it may not be surprising that the analytic potential of NAEP has not yet been fully exploited. It may well be worth undertaking a systematic exploration of what key analyses are desired from NAEP (or from a NAEP/SASS merger) to ascertain the degree to which -- and the ways in which -- they could be accommodated within the current structure of NAEP on a regularized basis.

I hope this is helpful to yours and the Center's efforts, Leigh. I thought the meeting was very useful.

Sincerely,

Linda Darling-Hammond
Director,
Education & Human Resources Program

LDH:nr

Stanford University
School of Education

MEMORANDUM November 29, 1987

TO: Professor Leigh Burstein

FR: Edward Haertel ed Haertel

RE: Reflections on the desirability of a NAEP/SASS merger

From the discussion at CES on Friday November 20, it seems clear that a NAEP/SASS merger at this time is ill-advised. Nonetheless, some of the ideas aired might lead to improvements in both NAEP and SASS.

What is meant by a NAEP/SASS merger? I take merging NAEP and SASS to mean that in 1990 or in 1992, the NAEP sample would be defined with schools rather than counties or county clusters as PSUs, and the same set of schools would then be asked to respond to the NAEP questionnaires as the SASS questionnaires, probably at about the same time.

Details of the sampling of respondents within schools under such a scheme are unclear. Presently, NAEP draws a sample of students and then administers a questionnaire to the teachers of those students sampled. In high schools, only teachers in particular subject areas are included, depending on the content area of the assessment. SASS draws a sample of all teachers in the school.

Threats to Continuity of NAEP Trend Data

Merging NAEP and SASS could jeopardize the continuity of NAEP trend data in two ways: by compromising school or teacher cooperation due to a more concentrated respondent burden and by altering the characteristics of the NAEP sample to accomodate SASS. Maintaining the NAEP trends must remain a paramount concern.

Concentrating respondent burden. In order to realize most of the potential benefits of a merger, it would be necessary to link SASS teacher survey responses to NAEP student data at least at the level of the classroom--linkage only at the school level would be much less useful. Thus, some coordination of NAEP and SASS sampling within schools would be required. This would concentrate the burden of responding on the teachers of NAEP student respondents, possibly leading to poorer teacher compliance. The increase in total person hours required for data collection in a sampled school could also jeopardize NAEP's exceptional school participation rate.

Altering the NAEP sample design. Further changes in the NAEP sample design and data collection procedures on top of those recently made to accomodate BIB spiraling and those already planned to enable state-level comparisons could also jeopardize the continuity of NAEP trend data. The magnitudes of biases that might be introduced by such design changes are difficult to estimate, but even small perturbations could disrupt trends.

Possible Justifications for a NAEP/SASS Merger

Despite these risks to the integrity of the NAEP program, several justifications might be offered for a NAEP/SASS merger. I do not find any of them compelling:

Merger would enhance the usefulness of NAEP, by tying measures of staff characteristics and fiscal variables to student achievement measures. Attempts to clarify what specific questions could be answered through such a linkage seem to end in one of two places. Most such questions could be answered through some modest redesign of the NAEP teacher and principal questionnaires. Presently, these questionnaires are driven by the student achievement data collection. If their purpose were conceived more broadly and more questions were repeated from one biennual survey to the next, the utility of NAEP might be enhanced.

The one sort of question that could not be addressed through small NAEP background questionnaire revisions concerns educational equity: How large are the disparities in quality and quantity of educational resources provided to different groups of learners? (E.g., children at risk, children in large urban school systems, children of different cultural and linguistic groups, the handicapped, and children well below average in achievement.) This is an important set of questions, but again, they would not be adequately addressed through a NAEP/SASS merger. Many aspects of these equity concerns could be addressed through SASS teacher survey questions about the characteristics of the students taught. An adequate accounting of total educational resources would be well beyond the capabilities of an enhanced NAEP, an enhanced SASS, or a merged NAEP/SASS data collection. Adequately addressing these concerns was not a primary objective of either survey. It might be accomplished through an intensive sample survey along the lines of Hall, et al.'s proposal, but is not a reasonable objective for NAEP and SASS.

Merger would enhance the usefulness of SASS, by providing measures of educational outcomes that could be tied to resource (input) variables. There is no question but that the usefulness of SASS as part of an education indicator system would be significantly enhanced by linkage to some broad student outcome measures. The issue is not just one of making the SASS database more attractive for secondary analysis, or pursuing academic research questions. Numbers from SASS could tell much more about the health of the

education system if it could reveal the educational consequences of different levels of staff qualifications or resource allocations. That being said, it does not follow that NAEP is a good source of the needed outcome information. Even though NAEP is now moving in the direction of providing summative achievement measures for individual students, its primary purpose remains to survey trends in aggregate performance on relatively narrow curriculum elements, and that is what it is designed to do best. I don't have a better solution. I am pessimistic about attempts to link or equate data from independent, ongoing assessments using different tests, and at the same time, I am reluctant to increase the testing burden on students. One possibility, especially in larger states, would be to link SASS data to state assessment data. California's CAP test, for example, provides solid data on ten percent of the students in the nation. If SASS instruments in California schools could be linked to CAP results, the usefulness of SASS could be increased without jeopardizing NAEP data.

Merger would save data collection costs and reduce total respondent burden. It appeared from Friday's discussions that there is not enough information available to estimate the magnitude of possible cost savings. Further study of this question would be helpful, but it is unlikely that savings would be sufficiently large to outweigh the risks of a merger to continuity of NAEP trends. It bears repeating that total respondent burden has less to do with respondent cooperation or with data quality than does the amount of time and effort required of individual respondents.

WESTAT
An Employee-Owned Research Corporation

MEMORANDUM

TO: Leigh Burstein

December 2, 1987

FROM: Morris H. Hansen, Westat MH

SUBJECT: CRESST NAEP-SASS Merger Discussions

I have looked over your summary of the meetings on November 18 and 20, and also Richard Murname's comments. I feel that I have little to add that hasn't all been covered in these two documents. They generally present a similar point of view with which I am in general agreement, subject to the following additional comments.

(1) Your point 7 in the summary states that serving national interests is an insufficient incentive to school (or district) participation, in the light of competing data collection burdens. This seems to suggest that specific feedback of individual school summaries into the schools or school districts from studies such as NAEP might be necessary to obtain cooperation. I believe that school benefits (and incentives) can be demonstrated through more general means, if the programs can be reasonably shown to be effective in guiding improvements in state and federal programs, curricula, etc., that of course benefit the schools. Effective cooperation with NAEP has been obtained in the past, without such specific feedback. I believe more extensive general uses and applications of NAEP and other worthwhile programs that are positive can be presented in a way to obtain cooperation, and should not be undersold. Other important national statistical programs in education and in other subject areas survive and have achieved effective cooperation without such specific feedback. Making cooperation depend on such feedback may lose the cooperation of schools that do not see an explicit benefit from the feedback.

(2) At least for the near future I believe it desirable to emphasize, as you have suggested, the desirability of fielding NAEP and SASS in different years (at least if NAEP is extended to a sample by states).

Again, your summary is not only an excellent summary of what was discussed, but presents a point of view with which I generally agree.

MH/jsn

cc: D. Bayless
A. Beaton

Comments on Merging SASS and NAEP

Richard M. Jaeger
University of North Carolina at Greensboro
6 December 1987

I agree with the developing consensus that NAEP and SASS not be merged in 1990. In making this recommendation, I am defining merger as:

1. Redefining the national NAEP sampling plan so that sampled teachers become a subset of those sampled for SASS,

2. Use of an expanded questionnaire for sampled teachers that incorporates virtually all questions presently used or planned for NAEP and all questions planned for SASS;

3. Use of identifiers that allow linking of student records, teacher records, school records, and school district records,

and

4. Ensuring that reasonably precise estimates of relational statistics can be formed for at least nationally representative samples of students in the NAEP-sampled grades and their teachers, students in the NAEP-sampled grades and their schools, teachers and their schools, teachers and their school districts, and schools and their districts.

I find the questions raised at the meetings on November 18th and 20th sufficiently compelling to convince me that the risks resulting from a 1990 merger of NAEP and SASS outweigh the potential benefits. In particular, the risk of jeopardizing the NAEP time series is substantial, and the nation can ill afford the disruption of that time series since NAEP currently provides the only trustworthy, nationally representative, longitudinal data on student achievement and academic progress. In addition, the potential benefits of a merger of NAEP and SASS, although discussed in the abstract in various CS documents, do not appear to be well articulated. And in the abstract, the case is not convincing.

The position advanced in the paper entitled *Alternatives for a National Data System on Elementary and Secondary Education* (Hall, Jaeger, Kearney & Wiley, 1985), that CS should develop an integrated national data system, rather than a series of unarticulated surveys, should,

in my view, guide the long-term redesign of the nation's program for collection of information concerning education and schooling. However, movement toward that goal should be gradual, based on a clearly articulated plan for analyzing and reporting resulting data, and based on a substantial body of research concerning the likely benefits and consequences of such movement.

Assuming postponement of a NAEP-SASS merger to 1992 or beyond, the intervening years should be devoted to the types of research necessary to more clearly guide a decision at that time. Much of the judgment concerning the possibility of merger in 1990 is based on speculation and essential caution, in the absence of clearly applicable information. In particular, resources should be devoted to:

1. Study of the effects of seeking information presently collected from teachers in NAEP and planned for SASS, on teachers' willingness and ability to provide such information. A carefully planned study could provide essential information on relationships between questionnaire length and content and teachers' response rates to the overall questionnaire,various types of questions, and various questions. Information that relates questionnaire length to data quality must also be obtained.

2. Study of the feasibility and costs of providing data-collection conditions for teachers that enhance response rates and the quality of data they provide, including alternatives to mailed questionnaires, payments to schools that would allow hiring of substitute teachers during data collection, and direct payments to teachers who provide data.

3. Detailed specification of the purposes to be served by a merger of SASS and NAEP, including a listing of the research questions to be addressed; the data series to be established or maintained; articulation of questionnaire items, data series, and research questions; and articulation of questionnaire items, research questions, and analytic procedures to be applied.

4. Beginning in 1990 at the latest, common record identification numbers should be used in NAEP and SASS, so that some data (however limited) from these surveys can be linked. Although such linking could not be expected to provide trustworthy national statistics, it would facilitate exploratory analyses that would illuminate the potential benefits of a formal merger of the two surveys. Record identification should allow both within-survey (vertical) and

between-survey (horizontal) linking of data. In addition to supporting exploratory analytic studies, such record identification would support estimation of the degree of respondent overlap and burden that results when both NAEP and SASS are conducted during the same year, and the comparative completeness and quality of data provided by teachers, schools and districts that are faced with one survey or both.

It is possible that school principals and superintendents (if not teachers) would agree to provide a substantial amount of data during a given academic year, provided they were assured that no federally-initiated data collection would take place within their schools (or school districts) in off years. Studies of the willingness of potential respondents to assume more intensive, but more widely spaced, periodic burden should be undertaken, as should studies of the potential advantages of using rotating panel designs for NAEP and SASS.

Illinois
State Board of
Education

100 North First Street
Springfield, Illinois 62777-0001
217/782-4321

EXCELLENT SCHOOLS: THE ILLINOIS ADVANTAGE

December 7, 1987

Dr. Leigh Burstein
Center for the Study of Evaluation
UCLA Graduate School of Education
405 Hilgard Avenue
Los Angeles, California 90024-1521

Dear Leigh:

With regard to the meeting concerning the possible merger of NAEP and SASS, the key analysis was made quite early by Linda Darling-Hammond. We need to keep reminding ourselves that the SASS Questionnaries are not blocks of granite but layers of pebbles that can be used when appropriate. Just because the questionnaires presently exist as a unit doesn't mean that they should remain so. It seems illogical to even maintain a dialogue about the eventual merging of NAEP and SASS as they presently exist. Your meeting convinced me that the approach is rather what components of each should be merged, when, and how often.

If the local responses aren't likely to vary much over a two year period, then why collect it? For this entire project to succeed, the response burden on the part of school staff and the analysis burden on the part of CES needs to be kept at a reasonable level. Therefore, your concept of "mild merger" is on target. Unfortunately, in my conversation with some CES staff who have presented these forms before CEIS, there have been instances when analysis questions were asked and the response was: Our job is to put the form together, get through the clearance process, and collect the data. Someone else will do the analysis. Although this is a paraphrase, I believe it reflects the present status. Even Emerson at our last meeting simply asked which items on which forms go with NAEP and which should be separate and how often should either be collected. It is not an easy task, but certainly manageable. I'm sure that a group that would include you, Linda Darling-Hammond, Dick Murmane and Paul Planchon could have a good product within a short time.

The first task would be to lay out the analytical framework of what questions need to be tied to pupil assessment data. Of those which need to be done biennially and which less frequently. In my chart I list the former with a "+' and the latter with "++".

Southern Illinois Regional Office
First Bank and Trust Building
Suite 214, 123 South 10th Street
Mt. Vernon, Illinois 62864-4013
618 242-1676

Telecommunications Device for the Deaf (TDD)
217/782-1900

State of Illinois Center
Suite 14-300
100 West Randolph
Chicago, Illinois 60601-3405
312/917-2220

Only when NAEP + and NAEP ++ has been decided should SASS be considered. Even the core of SASS shouldn't be collected sooner than every three years. If the questions and analysis procedures remain consistent, it seems unlikely that more frequent monitoring will be useful. This press to do it as often as possible is a legacy of the Congress and other publics receiving too many conflicting answers to the same question. Once the data base and the analysis process is credible, educators can spend more time reviewing successes and providing solutions to problems than dreaming up new ways to ask questions.

Assuming that the relevant school, principal, teacher and pupil achievement data via biennial NAEP (mild merger) is in place and that SASS is in place on a triennial basis, the question remains are there any reasons why these two data collections should ever occur during the same year. Perhaps there is a joint state or national profile that makes sense. I don't know, but there is time to investigate that possibility using this timeline.

	SASS	NAEP
1988	X	X
89		
90		X+
91	X	
92		X++
93		
94	X	X+
95		
96		X++
97	X	
98		X+
99		
2000	X	X++

The first NAEP + would be the mild, perhaps gentle, merger. SASS waits until 1991 when CES and others would have had an opportunity to carefully review the data and market the results. If a more expansive approach can be justified with NAEP ++, one can wait until 1992. It would not be until 1994 and every six years thereafter that they would occur during the same year -- the case still has to be made for the utility of doing that or perhaps the optional solution is to wait until 1991 and then move SASS to a four year cycle.

I could ramble on for a few more paragraphs, but I would simply start to repeat your comments and Dick Murmane's because they are so appropriate.

Thanks for the opportunity to comment.

Sincerely,

Thomas Kerins
Manager
Student Assessment Section

RMG 26291

December 3, 1987

Leigh Burstein
Center for the Study of Evaluation
Graduate School of Education
University of California
405 Hilgard Avenue
Los Angeles, CA 90024

Dear Leigh,

These rough notes are a response to your materials summarizing the NAEP/SASS merger meetings. I hope it arrives in time, and I apologize for its lateness and roughness. As you know, other events made it impossible for me to attend to this until a few days ago.

I strongly agree with your recommendation #1 that the NAEP and SASS not be mergeed at this time. I think, however, that the recommendations to the Advisory Council should clarify the reasons that various participants gave for avoiding the merger in more detail than is provided in your memo of 11/24. In particular:

- Valid analyses of school effects simply cannot be obtained from a national cross-sectional survey.
- Nonetheless, a merged NAEP/SASS would inevitably bring about a torrent of invalid but potentially influential studies of school effects that could seriously distort policy.
- Merger would seriously threaten the integrity of the NAEP as an indicator--that is, a descriptive study--of achievement. One reason is the risk of increased non-participation because of the increase in individual-level burden that merger would cause.

With respect to the first of these points, the limits and appropriate uses of cross-sectional data in general, and of nationally representative cross-sectional surveys in particular, need to be articulated more carefully before modifications are made to either NAEP or SASS, *even if full merger is ruled out*. There was pleasantly little opposition in the meetings to the strong position that Tony Bryk, Bill Schmidt, and I took about the limits of cross-sectional data--that is, that causal modelling of school effects is an entirely invalid use of such data. Nonetheless, the discussion of what uses are and are not appropriate was a bit unfocused, with a lot of alternative dichotomies (descriptive versus relational analysis, policy analytic versus school-effects research, etc.) being used without sufficient clarification. I would suggest the following elaboration.

First, causal modelling--or, more specifically, the testing of causal hypotheses--concerning the determinants of achievement simply is not an appropriate use of cross-sectional survey data. Such data can be used to *generate* causal hypotheses, but testing those hypotheses requires other types of data. The fact that they cannot be validly used to test causal hypotheses does not mean that cross-sectional surveys are unimportant. They can be an extremely valuable source of descriptive information. The NAEP, for example, is an invaluable component of our all too limited system of indicators of student achievement.

In the Wednesday meeting, you responded to this point by noting considerable disagreement about what is meant by "descriptive." The term is often used disparagingly, for example, when studies are called "only descriptive." Moreover, there is a widespread view that descriptive studies need to be technically simple, comprising bivariate cross-tabs and the like. In fact, neither view is warranted. Descriptive studies are simply those that attempt to figure out what a phenomenon is rather than to test a hypothesis about why that phenomenon came about. They play a critical role, in two ways:

- They shape further inquiry, by generating hypotheses and guiding other forms of research (such as smaller longitudinal studies designed to assess causal hypotheses); and
- They can provide valuable information for policy formation.

Moreover, descriptive studies can be technically complex. There is no reason, for example, why descriptive studies need to be only bi- or trivariate. Indeed, many multivariate studies that purport to be testing causal hypotheses are actually valuable because of the descriptive information they provide.

For example, Walberg and Fowler recently published a study in *ER* that argued that large school districts produce low levels of achievement, holding constant mean SES and per-pupil expenditures, and that expenditures are not significantly associated with achievement when district size and SES are controlled. The data were cross-sectional universe data for districts in New Jersey. In itself, this study cannot confirm or disconfirm the hypothesis that district size somehow causes the inefficient use of revenues, although it certainly makes that hypothesis more attractive. Nonetheless, it is valuable as a multivariate descriptive study, for it shows that certain important relationships hold (at least in New Jersey) even when conditioned on some important confounded variables.

If you accept this viewpoint, the purpose of relational analyses of databases such as the NAEP is to provide what could be called "conditional descriptive" information. For example, it is valuable (for policy as well as to guide other types of research) to explore the distribution of achievement, conditioned on ethnicity, region, type of school, and so on. These conditional descriptive analyses can of course be multivariate, subject to the limitations imposed by sample size and design and characteristics of the variables.

This then leaves us at the point where both the Wednesday and the Friday meetings came to a nearly dead end: what, precisely, are the conditional analyses of student achievement or school and teacher characteristics that we need both for policy and to guide other research? It is easy to come up with examples for the NAEP--that is, instances in which we need assessments of achievement conditioned on school, community, and other variables. Ethnicity is a good example. In addition to bivariate tabs and trend analyses conditioned on ethnicity (e.g., are blacks continuing to gain on whites?), it is important to consider a variety of trivariate relationships. For example, have the relative gains of black students been greater in high-minority or low-minority schools? In certain regions?

Examples where it would be productive to condition SASS analyses on achievement are less obvious (and probably far less numerous), but they exist. For example, it we want to track the flow of teachers with different characteristics into various types of classrooms within schools or schools within districts, the level of achievement of the students they are assigned is an obvious variable to include.

I think that the required next step is to rethink *systematically* what conditional descriptive analyses are important for both of the two purposes noted above, and to compare the results of that effort to the current variable lists for both the SASS and the NAEP. I think that the NAEP end of this should be relatively straightforward and might lead to the conclusion that the non-outcome variable set needs modification, perhaps by adding or substituting SASS items. The SASS end will prove far more difficult, for incorporating meaningful achievement measures into the SASS would be incomparably more difficult and more expensive than incorporating SASS items in the NAEP background variable set.

Give me a call if you would like to talk these issues over further.

Sincerely,

Daniel Koretz

November 10, 1987

Richard J. Murnane
Harvard University Graduate School of Education
Cambridge, Ma. 02138

Should NAEP and SASS Be Merged?

I recommend that NAEP and SASS not be merged at this time. I base this recommendation on an assessment of the probable benefits of the merger, and the possible costs. This memo sets out the reasons for my recommendation.

NAEP provides the most important information on the cognitive skill levels of American school children. Consequently, any change in design that threatens the ability of NAEP to provide unbiased information on achievement must be considered only if the risks of damage are low, and the potential benefits are high. In my view, the potential for damage in the form of noncompliance, or shoddy compliance by school personnel, particularly teachers, is significant. It took teachers about one hour to complete the SASS teacher questionnaire during the pre-test. If teachers are expected to complete this questionnaire carefully, and provide other information for NAEP about teaching techniques, this burden may simply be too great for a significant number of teachers. Moreover, it is likely that the teachers who do not provide complete cooperation will be teachers with particular characteristics, or teachers who work in particular types of school settings. Thus, such noncompliance, or less than complete cooperation, could jeopardize the sample design, and make it impossible to make valid inferences about the nation from the sample.

Another consideration is that NAEP is undergoing a change of its own -- moving to state-by-state comparisons. This change introduces a number of new issues concerning sample design and drawing inferences about the population from the samples. It seems unwise to me to attempt to introduce two major changes in NAEP at the same time.

The potential problems associated with merging SASS with NAEP are significant. In my assessment, the potential benefits are not commensurate with the potential problems. Let me consider three types of potential benefits in turn: increased analytical power, reduced respondent burden, savings on cost of administration.

Increased analytical power?

Merging SASS and NAEP will not enhance greatly the extent to which these data can support studies of school effects. One reason is that the cross-sectional nature of the NAEP design makes it inappropriate for causal modelling of school effects. For such causal modelling, longitudinal data on students' achievement are needed, such as are provided by HS&B, and NELS. A second reason is that the BIB spiraling used in administering the NAEP test items means that only a very few test items could

be mapped to the students of a particular teacher who completed the SASS teacher interview. Third, the information on teacher characteristics and teacher turnover that SASS will provide cannot be treated as "exogenous". Instead, the teacher characteristics and rate of turnover must be viewed as results of decisions teachers make about where they want to work and decisions school districts make about whom they want to employ. SASS may support research on the factors that influence teachers' and school districts' decisions. But it seems unlikely that models of this set of decisions could be combined reliably with estimates of the determinants of student achievement in one many-equation, sensible model.

Reduced Respondent Burden?

The primary question that CES should ask is whether the burden of participating in surveys will lead respondents to act in a manner that jeopardizes the usefulness of the survey information. Actions that threaten the survey include nonresponse, and careless completion of the questionnaires. Such actions are more likely the greater the burden that the surveys place on individual respondents. Consequently, while merging SASS and NAEP may reduce the total number of hours that teachers and administrators as a group spend completing questionnaires, it certainly will concentrate the burden on particular teachers and administrators. It is this concentration of burden that raises the likelihood of incomplete cooperation.

Reduced Cost of Administration?

Merging SASS with NAEP would probably reduce the cost of administering the two surveys by reducing the number of school districts that must be contacted, and visited. However, this source of cost savings takes place by increasing the respondent burden on personnel in the districts that would be selected for the joint NAEP-SASS survey. Thus, the savings in dollar cost are achieved by concentrating respondent burden, and thereby increasing the likelihood of less than full cooperation.

There is no question that issues of survey cost must be taken seriously, especially since the size of the CES budget is not known at this time. However, it seems extremely unwise for CES to take a course of action that threatens the quality of information provided by NAEP, the nation's best source of information on student achievement.

Questions about SASS

SASS is a promising, but untested strategy for learning about who teaches in the nation's schools. The pre-test results for SASS are very encouraging. Yet a pre-test does not provide nearly as much information about problems of administration and problems of interpretation as the first full fielding of the instruments will. An example of the many issues surrounding SASS concerns the demand and supply questionnaire. This is a lengthy and complex instrument. It took respondents several hours to complete this instrument during the pre-test. Only when the first administration has been completed and the data have been

analyzed will it be possible to assess reliably whether the amount of information provided by this instrument justifies the large respondent burden.

Given the many unresolved questions about SASS, it seems unwise for two reasons to combine it with NAEP at this time. First, the problems that may arise in fielding SASS could jeopardize the quality of information that NAEP provides about the achievement of the nation's school children. Second, it seems as if it would be easier to resolve the problems that may arise in fielding SASS if these problems are not complicated by issues of integration with NAEP.

In summary, this is the wrong time to merge SASS with NAEP. Altering the sample design for NAEP to accommodate state-by-state comparisons is a significant task. Ironing out the problems with the new instruments that are part of SASS will be a major task. Merging the two surveys at this time reduces the probability that the challenges that face NAEP and SASS will be met successfully.

UNITED STATES DEPARTMENT OF EDUCATION

OFFICE OF THE ASSISTANT SECRETARY
FOR EDUCATIONAL RESEARCH AND IMPROVEMENT

December 2, 1987

Leigh Burstein, Co-Director
CRESST Quality Indicators Study Group
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, CA 90024

Dear Leigh:

I appreciate being included in the NAEP/SASS merger meeting. While you did not specifically request feedback from in-house participants, my perceptions as a relatively naive newcomer may provide you with a perspective that you would not get otherwise. First I'll summarize what I heard. Then I'll summarize what I think.

What I Heard

There was general concensus that the separate, primary purposes of both NAEP and SASS are important. If (or however) a merger is implemented, the integrity of those separate purposes should not be compromised.

A recurring set of concerns focused on the relationships to be studied if a merger occurs. What relationships between and among variables will be examined and, more important, why? There also seemed to be some concern that any reported relationships might be misinterpretted as causal.

Another set of concerns focused on the burden of data collection on NAEP and/or SASS participants. The greatest fear seemed to be a potentially negative correlation between burden (actual or perceived) and validity of the data. There was also some concern regarding the payoff to schools for participation. Pertinent information in readily useable form was suggested as meaningful renumeration.

A number of participants suggested using existing NAEP data to inform any decision concerning a NAEP/SASS merger. Specific studies using NAEP data could focus a series of research questions about American

WASHINGTON D.C. 20208

schools that might, or might not, be efficiently/effectively addressed via some form of merger (e.g., strong, mild).

Participants repeatedly questioned which SASS components should be used in the event of merger. There was general agreement that not all aspects of SASS would need to be included in answering the questions addressed by the merger; hence, the continuing question -- what questions/relationships will any merger specifically address?

What I Think

NAEP was designed to provide national level data on student achievement. SASS was designed to provide estimates of school and staff characteristics that can be identified and aggregated at the state level. The primary rationale for a NAEP/SASS merger seems to be the ability to study relationships between student achievement and school/teacher variables that may be used to inform state, as well as national, level decision making.

I am convinced of the value (for the most part) of both NAEP and SASS purposes. If the intent of a merger is to examine relationshps between student achievement and what goes on in schools, I am not convinced that a NAEP/SASS merger (strong, mild, or otherwise) is an efficient, defensible approach. What relationships will be examined? Why? What relationships should be examined? Why? Are the "will be examined" and "should be examined" relationships the same? I fear not.

As a researcher and citizen I want to know why and how kids know what they know and what of that is attributable to particular teacher and/or school characteristics. Knowing that years of teaching experience and student test scores are positively related does not tell me why. In fact, I am not sure why studying that relationship, or any other (yet unspecified) relationship is important. I am concerned that the instrumentation (NAEP, SASS) will drive the questions asked and that the answers to those questions may have little meaningful impact on what happens to kids in schools.

Two important, related issues were not explicitly expressed in the meeting I attended (11/20/87). Those issues concern: (a) the

difficulty of uniquely attributing particular student achievements to particular sources (e.g., particular teachers) and (b) the fact that student achievement tests are designed to assess students' (limited) knowledge, not teaching or schooling effectiveness. I raise these issues because it makes intuitive sense that relationships between student achievement and other variables (e.g., teacher/school characteristics) indicate that: (a) particular student outcomes may be attributed to particular types of teachers or schools when, in fact, the variance not accounted for may be more informative than the variance shared and (b) student achievement test scores provide acceptable indicators of the effects of teaching and schooling when, in fact, such scores are but proxy measures.

I look forward to receiving your summary of the merger meetings!

Sincerely,

Doris

Doris Redfield, Ph.D.
OERI-CRESST Liaison

STATE OF SOUTH CAROLINA
DEPARTMENT OF EDUCATION
COLUMBIA 29201

Charlie G. Williams
State Superintendent of Education

November 19, 1987

Dr. Leigh Burstein, Co-Director
CRESST
Graduate School of Education
University of California, Los Angeles
405 Hilgard Avenue
Los Angeles, California 90024-1521

Dear Leigh:

This is in response to your request to attempt to capture the essence of the discussions of November 18, 1987 concerning the proposed merger of the SASS and NAEP samples. First, I believe there was strong concensus, if not unanimity, that a complete merger of NAEP and SASS is not desirable. The major points, as I recall them, in support of that position are as follows.

1. Interest in merging the samples seems to rest on the assumptions that NAEP and SASS should be conducted with the same frequency and concurrently. No compelling arguments have been advanced to support either of those assumptions.

   If the SASS and NAEP are to be conducted biennially, they should be scheduled in alternate years to spread data burden rather than concentrating the burden through merging the samples.

2. The proposed merger of the samples as a means of providing data for relational studies seems ill-advised.

   Although such studies may be informative and desirable, data bases of the magnitude of those generated from SASS and NAEP are not necessary for their conduct. The studies can be more effectively, and probably more efficiently, conducted with smaller samples and stricter controls than those provided by NAEP and SASS.

   In addition to the issues of efficiency and effectiveness, the following points were raised relative to relational studies: a) the strength of relationships between student achievement and other variables is not likely to change significantly in the short-term. Consequently, there is no necessity to collect data every two years to examine the relationships; b) the existence of large data bases linking teacher and school characteristics to student achievement may lead to inappropriate analyses and erroneous conclusions due to the temptation to apply a causal model to the interpretation of correlational studies; c) relational studies are most appropriately conducted on a longitudinal

basis rather than through cross-sectional surveys such as NAEP and SASS; and d) the data on teacher and school characteristics collected during the last two administrations of the NAEP have never been analyzed.

3. In the apparent absence of a model for the analysis and interpretation of the data on teacher supply and demand, the 1987-88 SASS data collection may not provide the information necessary to inform policy. The next cycle of SASS should not be scheduled until the data to be obtained from the 1987-88 survey have been analyzed and are available to influence the design.

4. Although the technical issues related to the merger of the samples can probably be satisfactorily resolved, the potential for negatively impacting participation rates in NAEP is too great to risk jeopardizing the assessments by merging the samples and concentrating the data burden. This is of special concern at this time since major changes are being proposed in the NAEP to provide state by state comparisons.

5. The primary goal of the NAEP is to describe what students know. That primary goal should not be jeopardized or subverted by burdening participants in the NAEP with supplying data which are, at best, tangentially related to NAEP's major goal.

Although the above comments certainly do not reflect all of the discussion concerning the issues, I hope that they at least capture the major points. If any of my comments need clarification, please contact me at 803-734-8258.

Sincerely,

Paul D. Sandifer, Director
Office of Research

/tmb

STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

SCHOOL OF EDUCATION
Office of the Dean

(415) 497-2111

November 20, 1987

Dr. Leigh Burstein
Center for the Study of Evaluation
UCLA Graduate School of Education
145 Moore Hall
405 Hilgard Avenue
Los Angles, CA 90024-1522

RE: NAEP/SASS Merges

Dear Leigh:

This will be short. I am convinced by the many and thoughtful issue papers that the merger idea has almost no redeeming value.

There appear to be two possible reasons for the merger -- that it would allow us to plow new research territory and that it would be less costly to carry out the merged surveys than it would be to do the two separately. Neither reason holds up.

1. Value to research: The long and fruitless history of attempts to relate teacher and staff characteristics and behavior as assessed by large scale survey instruments to cross sectionally gathered student outcome data should have convinced us long ago that it is only a mechanism for generating meaningless correlation coefficients. Our theory and our measurement sophistication are simply too weak to overcome the inherent difficulties in attempting to understand causal relationships with cross sectional survey data. Part 4 in the paper "Issues in the Combination of NAEP/SASS: Conceptual Issues" raises the proper issue in a carefully skeptical manner: "It will be necessary to determine the extent to which a cross-sectional data set would be an appropriate vehicle for investigating correlates of achievement,.....". The NELS88 and the earlier HSB longitudinal surveys are far better for such studies.

   The only research reason I can imagine for combining the surveys is to study the distributions of educational resources among various sub-groups in the population -- something like Chapter 2 in the 1966 EEO Report. This might be accomplished more simply by augmenting NAEP with a few carefully selected questions and perhaps with a school representative survey of teachers.

2. If everything went right, I suppose the cost might be reduced. The large number of sampling problems set out throughout the issue papers, however, are sufficient to convince me that there is a substantial chance of failure of the effort. There is a tremendous risk in putting all of the eggs in one weak basket. In light of the apparently very sensitive nature of the NAEP data collection the problem seems overwhelming. After all, at the present time we are not sure even of our capacity to carry out NAEP without a hitch. Multiplied by 50 to obtain state representative samples for NAEP the proposal to combine the surveys seems like sheer folly. If we were to set up a decision model my prior is that the probability of a partial or complete breakdown of the combined survey would approach 1.0 -- and there is a considerable chance that the breakdown could occur and not be identified for some time.

Best wishes,

MSS

Marshall S. Smith
Dean

POLICY STUDIES ASSOCIATES, INC.
1718 CONNECTICUT AVENUE, N.W. · SUITE 700 · WASHINGTON, D.C. 20009 · (202) 939-9760

December 3, 1987

MEMORANDUM

To: Leigh Burstein
From: Brenda J. Turnbull BJT
Subject: Merger of NAEP and SASS

Thank you for including me in the meeting you chaired for CES on issues in the possible merger of NAEP and SASS. The group raised a number of very important substantive points that will surely be of interest to CES's Advisroy Council. In this memo, I would like to emphasize a procedural suggestion that I think would help CES and the Council in continuing to clarify the issues.

I believe that the deliberations on merging NAEP and SASS should begin with a systematic analysis of the questions that CES would like to answer—in other words, the construction of an analytic plan for a hypothetical data set. The staff time invested in such a plan would greatly clarify both the potential benefits and the limits of addressing these questions with nationally representative (and state-representative) data. I think the limits would come into sharp focus and would provide good reasons not to undertake the merger, but if the merger did go forward then the planning effort would have laid important groundwork for the eventual data analyses.

Constructing an analysis plan would, for example, force CES staff to think through a model of the determinants of student achievement. Such a model has an important bearing on decisions about data collection, as our meeting made clear. A model of learning as a long-term process leads to this conclusion: cross-sectional data on the characteristics of teachers and schools can help in analyzing the factors that contribute to student progress over one year, but they will not be very useful in the absence of data about the students' beginning achievement levels, which were shaped by a multitude of factors at home and in previous schooling. By anticipating the types of causal statements that could conceivably emerge from the analysis of a merged NAEP and SASS effort, I think CES would find these statements would be so hedged with caveats as to be fairly useless.

An analysis plan could do other things as well:

- It could include many descriptive questions that a merged data set could answer. This would include questions about the types of students who receive instruction from teachers with particular backgrounds and qualifications. However, even in this area a good plan would consider the extent to which this data set would capture the important variation <u>within</u> schools.

- To the extent that CES wants to investigate causal relationships between schooling variables and achievement variables, the analysis plan should identify any such relationships that are best addressed with a nationally representative or state-representative snapshot, as opposed to smaller-scale or longitudinal studies.

- It would include consideration of how often the data need to be updated. As we discussed briefly at the meeting, trends of the sort that SASS will capture may not be so fast-changing that they require data points every two years, particularly with samples that are representative of every state. A three- or four-year cycle might be perfectly adequate. Collecting representative data for only a subset of states in each cycle might be another possibility, if the data could be weighted in such a way as to be nationally representative.

Another immediate step for CES would be to look at the data already in hand from teachers and administrators in the NAEP sample. What questions can these data answer? How do they need to be supplemented? Is a merger with SASS a way of supplementing them, or would smaller, more focused studies do the job better?

In summary, I think CES has begun to ask good questions about the wisdom of merging two large national efforts. Your summary of our meeting and the other written comments will give the Advisory Council a good set of arguments to ponder. My aim in this memo has been to suggest that good research management really has to work backwards—to begin with a set of questions one would like to answer, to construct analytic models that can answer the questions defensibly, and only then to plan the data collection that will fit the models. In this planning context, the considerable respondent burdens of national studies can be weighed and justified.

December 10, 1987

CRESST QUALITY INDICATORS STUDY GROUP

Report from Meetings on CES Merger of NAEP and SASS

ITEM VII

Center for Education Statistics NAEP/SASS Memoranda on Issues Sent to Meeting Participants

1. "Merger in the Combination of NAEP/SASS: Conceptual Issues"
2. "Issues in the Combination of NAEP/SASS: Global Sampling Issues"
3. "Issues in the Combination of NAEP/SASS: Analytical Sampling Issues"
4. "Issues in the Combination of NAEP/SASS: Respondent Burden"
5. "Issues in the Combination of NAEP/SASS: Management"
6. "Issues in the Combination of NAEP/SASS: Pilot Studies, Simulations, and Simple Tabulations"

September 16, 1987

Merger of NAEP and School/Staffing Surveys

The Center for Education Statistics is developing new data collection systems responsive to statistics needs of diverse users. Among other things, the Center is assessing the feasibility of a policy to begin combining, in 1990, the National Assessment of Educational Progress (NAEP) with the new School And Staffing Survey (SASS).

As the Center progresses through this exercise with NAEP and SASS, there are three goals it is trying to achieve:

1) Collection and maintenance of a unified data set that could relate specific policies, mixes of resources, and changes in the instructional system to outcomes;

2) Lessening of burden on schools, school districts, and teachers; and

3) Reduction of costs to the Federal Government for the collection of these data.

While the goals appear valid and desirable on their face, they raise questions of "why", "to what extent", and "how." Some questions, concerns and issues include the following:

1) How can the Center deal with the conceptual distinction between surveys with different purposes and divergent universes: (a) one sample of all schools with grades in range of K-12 for SASS and (b) three individual samples of U.S. schools for 4th and 8th grades and 12th grade for NAEP?

2) Is the assumed reduction in data burden by combining the surveys in 1990 really a shift in burden (fewer schools but more burden in each school)? Will schools actually perceive a huge increase in burden when they are included in the sample? And, if so, would the quality of responses be affected for any of the parties (i.e., administrators, NAEP teachers, other teachers, students)?

3) Following the data quality question, above, should participating schools be in rotating panels beginning in 1990 so studies of change can be enhanced or does the data burden issue demand that each data collection be from a fresh sample?

4) Year 1990 is intended to be a practical trial of a State representative NAEP (one course in one grade) together with merged data collection about schools and teachers considering that the remaining Schools and Staffing data will be collected by a separate contractor, what technical and management questions should be addressed (e.g., common instruments processed independently/or by one contractor for inclusion into the data base)?

5) Will the integration necessitate design changes that wi shift emphasis from the primary goals of each of the individual surveys?

6) Assuming a longer national survey and a shorter State survey of teachers and students outcomes, what would be the consequences of examining relational questions at the national level vs. on a State-by-State level (assuming that the sample for the NAEP portion is a State sample in 1990 or 1992)?

7) Should the cluster size in the teacher sample be increased to permit statements about the set of teachers in a school? The issue is one of being able to represent the set of teachers as a characteristic of a school, rather than having only a small cluster of teachers that would allow statements about teachers in general with no link to specific schools.

   Given that NAEP samples teachers of students to describe teaching methods and SASS samples teachers in schools to determine teacher characteristics, can these two goals be achieved with a common sample?

8) How can this merged system best be managed, given that it requires (a) test administration, (b) surveys to be completed by students, teachers and administrators, c) large scale data management and (d) both grantee managed NAEP and Federally managed SASS components.

Issues in the Combination of NAEP/SASS

Conceptual Issues

This issue paper deals with four sets of conceptual issues related to the merger and use of the data in a merged NAEP/SASS.

1. Compatability of Objectives

In 1978, in its continuing quest for comprehensive and dependable information on student achievement, in Section 405(k) of the GEPA, Congress specifically directed NAEP to carry out certain assessment activities:

- o collect and report at least once every five years data assessing the performance of students at various age or grade levels in each of the areas of reading, writing and mathematics;
- o report periodically data on changes in knowledge and skills of such students over a period of time;
- o conduct special assessment of other educational areas as the need for additional national information arises;
- o provide technical assistance to State educational agencies and to local educational agencies on the use of National Assessment objectives, primarily pertaining to the basic skills of reading, mathematics, and communication and on making comparisons of such assessment with the national profile and change data developed by National Assessment.

Historically, NAEP has collected some information on characteristics of respondents' communities, including the region of the country in which the community is located, its size, and socioeconomic status. NAEP has in addition measured a few student background variables, such as race and ethnicity, age, sex, and parents' educational attainments. The objective of this collection of background variables is to be able to translate them, together with the assessments, into meaningful guides to educational practitioners for the improvement of education.

The School and Staffing Survey has as its immediate objective to create a comprehensive data base that can be used to (1) profile the nation's elementary and secondary teaching force; (2) enhance assessments of teacher supply and demand by teaching field, level and location; and (3) examine school policies and practices, administrator characteristics, and teacher workplace conditions. The ultimate objective to which the SASS data contribute, along with other data acquired in CES surveys, is the discovery of those conditions, methods and practices that seem to make for better and more effective teaching and learning in the nation's schools and to make that information available to those who make policies for, and those who operate, the educational enterprise.

To achieve the objectives to which the School and Staffing Survey contributes, it is necessary to measure the effectiveness of teaching and learning in the nation's schools; i.e., to assess educational progress, and to be able to relate differences in effectiveness to the varying characteristics of teachers, administrators, schools, and the community. To have a separate, and even partially duplicative, student evaluation as a part of the SASS is unacceptable from the standpoints of cost to the government and burden on the schools, teachers and students. Therefore, CES must explore the questions of links between NAEP and SASS, including particularly costs and feasibility.

However great may be the compatibility of the objectives of NAEP and SASS, there remain great difficulties in making the process and procedures equally compatible, and there are some who have grave reservations that making NAEP data useful for a greater range of purposes will undermine the assessment's capacity to perform its basic mission effectively. There is concern that the dilution of resources and distortion of purposes can result from extensive use of NAEP for district or school building comparisons, or from efforts to link NAEP to other assessments or data collection efforts.

There are two very specific procedural considerations in combining NAEP and SASS samples. The NAEP sample of schools is limited to schools containing 4th, 8th, and 12th grades; the sample of teachers is derived from the sample of students within the schools. In contrast, the SASS school sample includes all schools, and the teacher sample is a probability sample of teachers in all grades. To accommodate these differences while maximizing the utility of the data acquired, it will be necessary to analyze the costs, burdens and benefits of a variety of sampling approaches.

Finally, there is the problem of linking one survey process that is deliberately insulated from Federal operation so that there will be no Federal test of students or Federal evaluation of teaching methods, and another survey process that is operated directly, or through a contractor, by the Federal government.

2. Potential Added Value of a Merger

The potential analytic advantage of merging SASS and NAEP is that the resulting dataset would contain more comprehensive information, and therefore would permit the investigation of more relational issues. There are two distinct ways in which this would come about: by increasing the information base at a given organizational level, and by permitting a new combination of organizational levels to be studied. The organizational levels of interest here including the student level, the teacher level, the school level, and the district level. The relational issues are primarily those of studying correlates of student educational achievement.

Increasing the information base at a given organizational level applies particularly to teacher information. NAEP currently permits student outcomes to be related to a small set of teacher variables, e.g., measures of special training. Merger with SASS would introduce additional teacher variables, such as:

- Teaching Status
- Teaching Experience
- Teaching Load

These additional variables could serve as potential predictors of student outcomes, either singly or through development of multivariate models.

The new combination of organizational levels that would result from a SASS/NAEP merger is the student/district combination. The merged dataset would allow study of district variables as predictors of student outcomes. This is not currently possible, since NAEP does not collect data at the district level and SASS does not collect data on student outcomes (except for overall graduation rates and college application rates). District characteristics that could be related to student achievement include:

- o Teacher Pay Scales
- o Graduation Requirements
- o Hiring and Retirement Policies

Again, the additional variables might be of interest as individual predictors or as components of multivariate models.

There are two basic questions that might be considered in this context. First, how valuable would the additional analytic capabilities resulting from the merger be? Second, to the extent that they are valuable, is it better to merge the two surveys or to simply augment NAEP to include more potential predictors of student achievement?

3. Types of Relationships to be Investigated

The preceeding issue — the potential added value of a merger — is somewhat abstract, in that it addresses the general value of relating student outcomes to variables measured at higher levels of aggregation. It is also necessary to consider the potential utility of studying specific relationships, and to decide whether combined SASS/NAEP data set is the best vehicle for this endeaver. Although this paper is not the appropriate place for setting out a list of specific relationships that might be studied, it does seem valuable to consider a dichotomization of relationships into those that are established and those that are hypothetical.

An established relationship, e.g., the effect of instructional time on achievement levels, could be addressed in two ways: It could be further confirmed, or it could be refined and studied in finer detail. Further confirmation would entail extending the results of case studies or of relatively limited surveys to a national population. Refinement would involve, for instance, establishing the differential effect of instructional time on different sub-populations, e.g., on different ethnic groups or in different regions of the country.

A case can be made, however, for not conducting this type of research. The alternative would be to accept an established relationship as given, and to simply measure the indicator, i.e., the correlate of achievement. If this approach were taken, then the case for merging SASS and NAEP would be less strong.

Alternatively, the combined dataset could be viewed more in terms of exploratory analysis, i.e., as a tool for formulating and testing new relationships. Although new relationships do not necessarily imply new data

elements, they tend to do so, and the exploratory approach could well lead to lengthier survey instruments. This could lead to valuable research results. On the other hand, it might be more effective to conduct this research through special studies, rather than appending it to a major national survey.

4. Utility of Cross-Sectional Data

When investigating the correlates of educational achievement, it must be recognized that current achievement level is not simply a function of the current educational environment. It is, rather, a cumulative function of educational inputs that started in kindergarten or earlier.

Longitudinal studies, e.g., NELS, can measure educational inputs over a period of years, and attempt to develop models that predict or explain variation in educational attainment. Alternatively, studies that include pretests and posttests can measure changes in educational attainment, and relate these to current inputs.

Both SASS and NAEP are cross-sectional studies, and will remain such, whether they are combined or kept separate. It will be necessary to determine the extent to which a cross-sectional dataset would be an appropriate vehicle for investigating correlates of achievement, and to consider enhancements that might make the dataset more appropriate.

ISSUES IN THE COMBINATION OF NAEP AND SASS

Global Sampling Issues

D. The primary issue to be resolved is that the samples for NAEP and SASS were originally designed for two different purposes. There is some concern regarding combining the two surveys, since the final sample design for the combined surveys would necessarily be a compromise which may satisfy neither set of goals. Items I. - IV. below describe the essential differences between the two survey sample designs, and item V. describes what compromises look most appealing at this time.

I. Question 1 : How can the Center deal with the conceptual distinction between surveys with different purposes and divergent universes?

Issue: NAEP and SASS currently use different sampling frames. A sample design that would be used for both surveys must meet the needs of both surveys. Since the universes are different, this means that we would like to maximize the overlap between the two frames and samples, and use stratification to define relevant sets to use in estimation.

NAEP studies three universes:

1. the set of all schools which have a grade four;
2. the set of all schools which have a grade eight; and
3. the set of all schools which have a grade twelve.

whereas SASS studies one universe: the set of all schools which have any grade in the range K-12 inclusive. The sample design must accomodate both (or all) universes to allow estimation for the entire U.S., while at the same time allowing the time series established for NAEP to continue for each of the three universes. (Schools which fall into at least one of the three NAEP universes comprise 96.2 percent of all schools in the U.S. which have a grade in the range K-12.

Question Numbers refer to numbers used on the document: Merger of NAEP and School/Staffing Surveys

II. Question 6: Assuming a longer national survey and a shorter State survey, what would be the consequences of examing relational questions at the national level vs. on a State-by-State level?

Issue: NAEP and SASS currently provide estimates at different levels of aggregation. The design of the sample for SASS allowed for State-by-State comparisons for schools and teachers, whereas the other comparisons to be made from the survey (e.g. public vs. private schools and teachers) were only incorporated into the design assuming national level comparisons. The issue here is really the importance of the relational questions relative to other goals from NAEP and SASS.

NAEP for 1990 is intended to be:

1. Nationally representative for grades four and eight;
2. State representative for the assessment of progress in mathematics for grade twelve; and
3. Nationally representative for all other assessments in grade twelve.

SASS for 1990 is designed to provide:

1. National estimates for characteristics of schools
2. National estimates for characteristics of teachers
3. State comparisons for characteristics of schools
4. State comparisons for characteristics of teachers
5. National level comparisons between public and private schools
6. National level comparisons between public and private school teachers
7. National level comparisons between elementary and secondary schools
8. National level comparisons between elementary and secondary school teachers
9. National level comparisons between fields taught for secondary school teachers

SASS can also provide national level comparisons; for example, it can be used to make comparisons of large vs. small schools or teachers, or for urban vs. rural schools or teachers, but the sample design for the 1988 survey did not explicitly account for these comparisons.

III. Question 5: Will the integration necessitate design changes that will shift emphasis from the primary goals of each of the individual surveys?

Issue: NAEP and SASS currently provide estimates for different substantive populations. NAEP provides estimates of:

1. the assessment of progress for students;
2. the characteristics of teachers as they relate to progress.

NAEP can provide estimates at the national level for school or school characteristics, and at some levels below national (e.g. regional, urban/rural), but the sample is not well balanced across states for the school estimates. This is due in part to the emphasis on teachers and students, and in part because of the clustered nature of the sample, where counties are used as the first stage units.

SASS provides estimates of:

1. characteristics of teachers, and
2. characteristics of schools and school districts.

An integrated survey would attempt to optimize the sample so as to provide the best estimates for all of these goals, while at the same time considering some of the relational issues. The last issue (II. above) focused on the relative importance of the level of aggregation. This issue is more concerned with the relative importance of the variables being studied at the same level of aggregation.

IV. Question 3: Following the data quality question, above, should participating schools be in rotating panels beginning in 1990 so studies of change can be enhanced or does the data burden issue demand that each data collection be from a fresh sample?

Issue: NAEP and SASS are both recurring surveys, but neither of the current sample designs take account of the possible efficiencies of a rotation design. Both designs call for unduplication between the two surveys and NELS:88 in 1988.

The NAEP sample design selects counties or groups of counties as the first stage of selection, with schools at the second stage clustered within counties (initially thought to keep test costs down, though this point is under contention now). SASS is designed selecting schools as PSU's from a list, with an area frame supplementation for private schools. If a rotating design were to be implemented, to meet the objectives of both surveys, the rotation could occur at either of two levels: the county level and the school level. Determination of the design for the combined surveys will be a function of the costs and size of the survey.

If the combined survey were a state sample for all grades, it may be that there would be enough schools in sample that the cost savings realized from the elimination of between county travel would vanish. If the combined survey were only a national sample for parts of NAEP, it may be that savings would be substantial for NAEP to start with a sample of counties as PSU's. However, the advantages of a rotating design are reduced for NAEP because different topics are assessed each time; the recurrence of topics is staggered.

V. Current proposals: some ideas on a combined design.

The sampling frame for the combined surveys would be all schools with any grade in the K-12 range. Schools would be allocated to multivariate strata, with one of the stratification variables being whether a school has a grade 4, 8, or 12, some combination of 4, 8, or 12, or none of these. Estimates would be made for SASS from the entire sample. Estimates would be made for NAEP for grades 4, 8, and 12 using only the appropriate strata.

Determination of the number of schools and teachers to be sampled will be a function of several factors:

1. The costs of interviewing schools, teachers, and students.
2. The types of analysis to be conducted using schools, teachers, and students, and the relative importance of each of these analyses.

For 1990, the sample for the combined survey should be a national sample with state supplementation for the portions of SASS and NAEP that will require state estimates.

The determination of whether a rotating design should be used will be a function of:

1. Whether the analysis of data from the combined sample will have a component related to school context.
2. Whether there may be a problem with burden if schools are sampled repeatedly. This may be a nonissue if large schools will fall into sample with certainty for a state sample.

Covan/TODO/NAEPSSS1.ISS

ISSUES IN THE COMBINATION OF NAEP AND SASS

Analytical Sampling Issues

O. The most important issues for NAEP and SASS related to the use of the data are the production and maintenance of time series.

- For NAEP, a time series using the universes of schools with grades 4, 8, and 12 for national estimates of student performance in various subject areas are of paramount importance.

- For SASS, development of a time series nationally is as important as the primary cross-sectional goals: school and teacher estimates for the public/private sector, for the elementary/secondary sector, for states, and for secondary school teacher estimates by field taught.

These primary goals are somewhat in conflict with each other, and individually could lead to different sample designs. The primary sampling question is how the analysis plan can be used to determine how to develop the sample design.

I. Question 5: Will the integration necessitate design changes that will shift emphasis from the primary goals of each of the individual surveys? and
Question 7, part 2: Given that NAEP samples teachers of students to describe teaching methods and SASS samples teachers in schools to determine teacher characteristics, can these two goals be achieved with a common sample?

Issue: What is the importance of the relational analysis relative to the primary goals of the individual surveys? and,

Issue: In establishing a model for the relational analysis, one must consider that there are different and varying influences to consider. For some portion of the sample, a class or subsample of students will have only one teacher (e.g. fourth graders), and so some inferences can be made involving specific teachers tied to clusters of students. For another portion of the sample, a subsample of students in a specific grade will have several teachers, and the degree of overlap between teachers and students in a school will be very fuzzy (e.g. 12th graders). Finally, some students in a school may have just transferred in, whereas other students may have gone through several grades in the same school where they are now sampled.

*Question Numbers refer to numbers used on the document: Merger of NAEP and School/Staffing Surveys

"What teachers to sample" is a function of what influences are to be considered in the model. How much can be considered realistically in a model? What is practical to collect? What does one do to reflect all of these influences? Or is the issue one of selecting only certain influences to be included in the model?

The relational analysis can be conducted at several levels. In a model of inputs related to outcomes, several sectors may be important factors:

1. Effects of specific current teachers;
2. Effects of past teachers, represented as a set;
3. Effects of a particular school in terms of environment;
4. Effects of a particular school district in terms of characteristics of the local population;
5. Subject matters covered in the past and present;
6. Instructional practices used in the past and present;
7. Other factors (e.g. demographics) which are needed as controls.

It may also be important to consider characterisitcs of other students in classes in which the sampled students are located, as another set of environmental effects. A third class of factors might relate to parents and other non-school or non-teacher related items.

Some sampling decisions are related to the planned analyses. Should a large sample of schools be taken within a county to provide estimates of school district environment. Should a small sample of schools be taken within a county to optimize the school estimates nationally and by state? Should a large sample of teachers be taken within a school, both to relate teachers to students and also to provide a measure of school environment? Should a small sample of teachers be taken within a school to optimize the teacher estimates nationally and by state?

There are also some issues of trying to oversample certain subpopulations to make comparisons. We can oversample schools in certain types of school districts or counties to ensure large minority representation for comparisons. We can oversample teachers by area taught or by characteristics identified in a screening interview. We can oversample students, and ultimately parents, again after a screening interview, to represent minorities or other factors important to a relational analysis. The decision regarding oversampling of minorities is entirely a function of where it is most important to the analysis to have comparisons of minorities to the balance of the sample. If this is only important to the relational analysis, the oversampling occurs at the last stage. Other comparisons may demand oversampling of minorities at an earlier stage.

Issues in the Combination of NAEP and SASS

Respondent Burden

One of the stated objectives for combining NAEP and SASS is to reduce the respondent burden. By using a common sample of schools for both surveys, a significant reduction of burden at the school level may be achieved.

The 1988 SASS will sample approximately 13,000 public and private schools, 65,000 teachers, and 5,600 public school districts. If the recommendationsof the Alexander/James Study group are implemented, the NAEP sample would increase in size to include 700,000 students, and approximately 14,000 schools and 60,000 teachers. A common sample of schools could reduce the number of schools by a factor of 1.5 to 2. However, several additional points need to be cited.

- The assumed reduction in burden by combining the surveys may really constitute just a shift in burden from many schools with relatively light burden to fewer schools with substantially increased burden. If the schools perceive this as increased burden, will the quality of responses to lowered?

- Reducing the number of schools in this way is unrelated to teacher burden. To the extent that the teacher samples for the two surveys are non-overlapping, teacher burden will not be substantially reduced.

- If a combined teacher sample is used, the burden on the individual teacher who is responding to both the NAEP and SASS data requests will increase significantly, perhaps by as much as 50 percent.

- At some level, burden may become so large that we lose the cooperation of our data providers. The 1987-88 SASS provides some insights into this potential problem. 1) The sample in some small states exceeds 50 percent of all schools. 2) In five large school districts, more than 50 schools have been selected, and in New York City 190 were selected. The Center may also anticipate that some states may choose not to participate in an expanded State-representative NAEP when the national student sample reaches 700,000.

Other approaches to controlling burden include:

- Control sample selection at the school level to ensure that a school is only included every other survey cycle or every third cycle.

- Incorporate matrix sampling into the questionnaire design for SASS. However, this would limit the usefulness of SASS for relational analysis.

- Reduce the questionnaire content and target questions at relatively narrow topics of interest.

Agenda - Item D

Issues in the Combination of NAEP and SASS

- Management

I. NAEP and SASS are operated under different management structures.

A. NAEP is operated as a grant with an external governance structure.

1. Funding for NAEP comes under a special NAEP line in the Education Department budget with authorization and appropriation set by Congress.

2. Grant awards are made through a competitive process, currently defined by general Department regulations and in the future by regulations specific to NAEP.

3. Decisions about the design and policies of NAEP are, by statute, made by the NAEP governing board, the Assessment Policy Committee (APC) and its subcommittees - e.g., Learning Area Committees, the Background Review Committee, and the Technical Advisory Committee. The OERI Assistant Secretary is an ex officio member of the APC.

4. Project activities are carried out by the grantee, currently Educational Testing Service, and its sub-contractors - i.e., Westat has responsibility for sampling and field operations.

5. Analyses and reports (including publication approval and dissemination) are done by the grantee and, to a lesser extent, secondary analysts.

6. The grantee develops a clearance package which is reviewed by CES, FEDAC, and OMB.

7. Additional NAEP related planning activities are being conducted by a Consortium on Assessment, organized by the Council of Chief State School Officers and funded by ED and NSF.

B. SASS is operated as an interagency transfer/agreement with the Bureau of the Census.

1. Funding for SASS is part of the general funding for CES in the ED budget.

2. Decisions about the design and policies of SASS are made by the Education Department.

3. Planning activities have been carried out by CES, by Rand under contract to CES and by the Bureau of the Census.

4. Sampling has been done by CES and the Bureau of the Census.

5. Field operations for 1988 will be conducted by the Bureau of the Census.

6. Analyses and Reports for 1988 SASS data will be done by CES and under competitive contract.

7. Publication approval and dissemination are determined by ED.

II. NAEP and SASS have certain goals in common.

A. Provide timely, useful information to a variety of audiences.

B. Control operating costs.

C. Control respondent burden.

D. Be responsive to the interests and concerns of:

- o data providers at State and local levels, and
- o the Education Department, the U.S. Congress, OMB, etc.

III. NAEP and SASS interests partially overlap both in content and respondents

A. School Questionnaire: Most of the items on the NAEP school questionnaire also appear on the SASS school questionnaire. The SASS school instrument is larger primarily because of additional questions about (1) the school administrator and (2) teacher supply and demand.

B. Teacher Questionnaire: There is less overlap in the teacher questionnaire. Both contain questions about teacher background/preparation and perceptions of school policies/practices but they differ in emphasis. They are in fact quite complementary.

1. NAEP's focus is on factors most related to assessment outcomes, especially the classroom and instructional practices in the subject matter area of the assessment. (NAEP teacher sample consists of each student's current teacher in the assessment subject.)

2. SASS provides more depth on teacher training, experience and attitudes. (SASS teacher sample consists of a random sample of all teachers in in the school.)

C. Student Assessment Instruments (NAEP only)

IV. Coordination of NAEP, SASS, and NELS for 1988

A. CES reviewed school and teacher questionnaires from NAEP, SASS, and NELS:88 and recommended changes for increased consistency of related items. All parties agreed to make the changes: ETS, Westat, Bureau of the Census, and NORC.

B. CES recommended a target of zero school overlap in samples for the three surveys in 1988. All parties agreed and cooperated. Westat/ETS drew the NAEP sample in June; NORC drew the NELS:88 sample in July based in part on Westat/ETS information; and CES and the Bureau of the Census are drawing the SASS samples based in part on NAEP and NELS information provided by NORC. (SASS public school sample was completed in August but the SASS private school sample has not yet been completed.)

C. CES sent two letters to Chief State School Officers, a June letter describing our plans for coordinating the surveys and an August letter reporting near zero public school overlap nationally and specific sample information for the State.

V. Coordination of NAEP and SASS for 1990

A. CES staff to develop milestones for coordinated planning and implementation of 1990 surveys.

B. CES staff to develop (1) analytic agenda and (2) model for integrated samples for NAEP and SASS with both input and review by outside people.

C. NAEP items to be developed with NAEP grantee and SASS items to be developed by CES and Bureau of the Census.

1. NAEP items will be strongly influenced by the Consortium of Assessment and other scholarly/field input.

2. SASS items will be selected by ED in light of scholarly/field input.

3. CES will be actively involved in the coordination of these two developmental processes.

   a. CES to negotiate commitment to coordination with 30 month grantee

   b. CES to review all instruments to provide coordination across common items

   c. CES to monitor instrument development and convene meetings to iron out problems

4. CES work with field coordinated across NAEP and SASS projects.

   a. State coordinators in field collections

   b. CEIS

   c. Field input for design and instrumentation

   d. State cooperative program

D. NAEP grantee cooperation and effort needed to collect SASS items/instruments in overlapping schools.

   1. Design issues in overlapping schools

      a. School Questionnaire: one instrument or two? bridges (bench marks)?

      b. Teacher Questionnaire: one instrument or two? bridges (bench marks)?

      c. Teacher samples within schools

         1. school contact
         2. possible overlap of NAEP and SASS teacher samples

   2. Data sharing agreements

E. NAEP grantee and Bureau of the Census cooperation and effort needed to draw and implement overlapping samples. (See V C 3, above.)

VI. Management structure issues

A. Congressional action could change NAEP governance structure — uncertain.

B. Budget levels will affect sample sizes, analysis efforts, trade-offs, etc.

C. Field response could affect target samples as well as response rates.

D. Field advice: Don't put two different data collections in any school. If information is needed for two surveys from any single school, then be sure to fully integrate school contact and data collection in that school.

Item VIII

December 10, 1987

CRESST QUALITY INDICATORS STUDY GROUP

Report from Meetings on CES Merger of NAEP and SASS

ITEM VIII

Letter to Participants Regarding Summary and Recommendations from Meeting

UNIVERSITY OF CALIFORNIA, LOS ANGELES

UCLA

BERKELEY · DAVIS · IRVINE · LOS ANGELES · RIVERSIDE · SAN DIEGO · SAN FRANCISCO

SANTA BARBARA · SANTA CRUZ

CENTER FOR THE STUDY OF EVALUATION
CENTER FOR RESEARCH ON EVALUATION, STANDARDS AND STUDENT TESTING
UCLA GRADUATE SCHOOL OF EDUCATION
405 HILGARD AVENUE
LOS ANGELES, CALIFORNIA 90024-1521
(213) 825-4711
(213) 206-1532

December 10, 1987

TO: Members, Advisory Council on Education Statistics (ACES), CES
Emerson J. Elliott, Director, CES

FROM: Leigh Burstein *Leigh Burstein*

RE: Summary and Recommendations from Meetings on Merging NAEP and SASS, November 18 and 20, 1987

The enclosed materials constitute my report to ACES from the November 18 and 20, 1987 meetings organized by the CRESST Quality Indicators Study Group on merging NAEP and SASS. The packet includes a summary and recommendations based on the discussions at the two meetings and on the written statements provided by meeting participants and other invited consultants. In addition, selected materials provided to participants prior to and during the meeting, lists of meeting participants, and the full texts of statements provided by participants and discussants. I apologize for the amount of material; however, it is my understanding that ACES members have had differential exposure to the questions and issues that led to merger discussions. Therefore, I decided to be inclusive, thereby allowing the audience the discretion in judging their information needs.

To expedite consideration of the essential questions addressed by this activity and the recommendations it generated, a statement of the background of the meeting and discussion of the primary recommendations follows in this cover memorandum. The CRESST activity was in response to conflicting advice received by the Director of CES. ACES had previously recommended that a merger of NAEP and SASS proceed. This recommendation was in keeping with the recommendations on linking data collections from the report on alternatives for a national data system on elementary and secondary education prepared by Hall, Jaeger, Kearney and Wiley (December 20, 1985). Yet other segments of the educational community questioned the advisability of the merger on a variety of technical, substantive, practical, and political grounds.

The purpose of the meetings was to bring together persons knowledgeable about educational research, statistical, and policy analytic issues that CES's data collections (including NAEP, SASS, Longitudinal Studies) are intended to address to:

a. Consider the range of issues that CES had already identified and review its available documentation regarding these issues;

b. Augment CES's prior analyses with other evidence that bears on the perceived benefits and costs of the proposed merger;

c. Assess the likely consequences (e.g., for knowledge production, enhancing policy analysis capabilities, improving or degrading the quality of data) of the merger;

d. Recommend options with regard to the decision process on the possible merger and the steps that should be undertaken in advance of a final determination to proceed with the merger.

Participants were provided in advance specific questions and issues that the meeting was intended to consider and a set of pertinent documents. Two 5-1/2 hour meetings were scheduled with a day in between to accomodate the schedules of the desired participants and to allow time to prepare information from the first day's discussion assist the second day's deliberations.

Without going into detail, despite the diversity of perspectives and interests represented in both days' meetings, there was consistency in the basic issues that needed to be addressed and considerable consensus about the primary recommendations. Briefly, the list of issues is as follows:

1. What does "merger" mean and how comprehensive (with respect to instrumentation and to samples) should it be?

2. What analytical purposes should guide any merger decisions?

3. What are the likely consequences of alternatives with respect to respondent burden and costs?

4. How does the question of the desirable/necessary cycle/periodicity and timing of SASS (or parts of SASS) interact with the above?

5. What sets of analytical exercises/special studies should be undertaken to address the merger issue in both the short run and the long run?

The recommendations that achieved a general consensus from the meetings and written statements are:

1. A major merger of the questionnaires and samples from NAEP and SASS should NOT be attempted in 1990. The risks of overburdening NAEP in 1990 are too great; Moreover, too little is

known about how SASS will actually function at this time to assess the benefits and consequences of strong ties with NAEP.

2. Whether NAEP and SASS should merge in 1992 or 1994 warrants further study including analyses of existing data from the two surveys gathered through the 1988 data collection.

3. Regardless of the extensiveness of the eventual merger, the analytical purposes that should guide merger efforts should be those dealing with informing the policy analytic process rather than enhancing capabilities to conduct school effects or effectiveness research in an integrated national or state-representative data base. An example of a policy analytic purposes that could be served through a "merger" effort are the gathering and maintenance of national (and perhaps state representative) indicator series dealing with questions of access and participation (e.g., which kinds of students receive instruction in which kinds of schools from which kinds of teachers?)

4. For the short term (e.g., 1990), a small set of teaching and schooling conditions questions selected from SASS could be administered with NAEP to enhance its ability to serve policy analytic purposes. To this end analytical work using past NAEP collections of teacher and school characteristics as well as other efforts to identify specific policy analytic purposes to be served should be carried out in time to modify and augment the 1990 NAEP school and teacher characteristics questionnaires.

5. A three-year or even a four-year cycle for the major SASS data collection should be considered with at least part of the resource savings shifted to conducting special studies (e.g., longer term study of flow of teachers into and out of the workforce for a panel of schools and districts; augmentation of NAEP data collection in 1990; studies of the consequences of the intensity of respondent burden and costs consequences of major merger). Alternatively, the SASS instrumentation can be broken up into smaller sets which could be fielded on different cycles with perhaps a core set maintained on a more frequent cycle. Spreading out the SASS cycle would also postpone collection activities in ways that would place less strain on plans for the 1990 NAEP.

6. Postponing major merger discussions beyond 1990 provides time and resources to consider (through design and other special studies) the costs and benefits of developing a merged sampling universe across the major data collections (including NELS as well as NAEP and SAS).

7. Attention is needed to the benefits accrued at the school level from participating in these surveys. "Contributing to national well-being" is increasingly losing out given the extensiveness of data collection demands and competition from data collection with greater extrinsic rewards.

The above conveys only the tenor of the discussions and written statements. Participants seemed genuinely concerned that the primary purposes of NAEP and SASS not be sacrificed or damaged by a hurried decision to merge the two. CES is undertaking major modifications and extensions of its data collection responsibilities over the next few years. Under such circumstances the participants seemed to feel that time devoted to fielding and reporting these collection efforts in an effective and credible manner is critical. Discussions of mergers of these data collections need to proceed at a more deliberative pace than at present. There is just too much at stake.

I hope that you find the enclosed materials informative. I look forward to meeting with you to clarify and discuss any aspects of the meetings, documents, and issues.